Essays in Economic Theory

Dissertation
submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wirtschaftswissenschaften, Dr. oec.
(corresponds to Doctor of Philosophy, PhD)

presented by

Jean-Michel Benkert
from Sumiswald, Bern

approved in July 2017 at the request of

Prof. Dr. Nick Netzer
Prof. Dr. Georg Nöldeke

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 19.07.2017

Chairman of the Doctoral Board: Prof. Dr. Steven Ongena

Acknowledgements

In the past, I have compared my PhD studies with taking a ride on a roller coaster, as there are plenty of highs and lows along the way. Just as any good roller coaster, my studies brought me plenty of joy and excitement as well as feelings of unease and anxiety. Luckily, I was not riding on my own, so the highs were magnified while the lows were softened.

My advisors, Nick Netzer and Georg Nöldeke, played a huge role throughout my PhD studies. In fact, Georg got me onto the roller coaster in the first place by rousing my interest for economic theory during my undergraduate studies in Basel and getting me out of my comfort zone to pursue my graduate studies in Barcelona and Zurich. Once in Zurich, Nick took me under his wing and introduced me to the world of behavioral theory, which has captivated me and inspired most of my thesis. Most importantly, both of them showed me that while research requires plenty of work, it can be pursued in an environment of kindness filled with fun and laughter, thereby making you forget you are working at all.

I benefited tremendously from all of my colleagues in Barcelona, Basel, Chicago, and Zurich. They helped me enjoy my studies and improve my thesis. In particular I would like to thank Stefanie Bossard, Mirjam Britschgi, Lea Cassar, Samuel Häfner, Andreas Hefti, Vanessa Herzog, Priit Jeenas, Wischiro Keo, Arnd Heinrich Klein, Igor Letina, Shuo Liu, Chloé Michel, Eva Ranehill, Anne-Katrin Roesler, Florian Schaffner, Armin Schmutzler, Andreas Thomann, Timo Tondelli, and Samuel Verbeke.

I thank the UBS International Center of Economics in Society at the University of Zurich and in partiuclar Sally Peggs and Roman Studer for the generous support. The UBS Center Scholarship has allowed me to pursue my PhD studies with every freedom I could hope for.

As with everything, I could not have done it without my family. I thank my parents Claire and Erwin, my sisters Davina and Nathalie, my brother(-in-law) Matthias, and, in particular, my girlfriend and soon-to-be wife Jessica – j.v.a.

Finally, I would like to thank all my co-authors – it was a blast!

Jean-Michel Benkert, Zurich, March 2017

Contents

Part I

Dissertation Overview

Dissertation Overview

This dissertation consists of three separate chapters, all of which are focused on economic theory. While the chapters are only loosely related, there are some common themes. In what follows, I will give a brief outline of each chapter and try to highlight the connections between them.

In Chapter 1 I consider the problem of bilateral trade under the assumption that the trading agents are loss averse. The bilateral trade problem is a cornerstone in the mechanism design literature and goes back to seminal work by Myerson and Satterthwaite (1983). They showed that it is in general impossible to realize all the gains from trade between a buyer and a seller who are privately informed about their valuation of some good. The result is obtained under the assumption of quasi-linear utility. While standard, this assumption does not allow to address some phenomena such as the endowment and the attachment effect, which have been documented empirically in trade situations. The former refers to the finding that sellers value a good more simply because they own it and the latter to the finding that buyers value a good more when they expect to buy it. Expectations-based loss aversion offers a way of explaining these effects, thus providing the motivation for the introduction of loss aversion into an otherwise standard bilateral trade setting in Chapter 1. I find that the presence of loss aversion can mitigate the severity of the impossibility result but not reverse it. I further consider the problems of maximizing the designer's revenue when acting as a broker as well as the problems of maximizing material and total gains from trade. I find that the optimal mechanisms insure agents against ex-post variations in their payoffs by entailing interim-deterministic transfers and by reducing the trade frequency. Interestingly, the mechanisms maximizing the material and the total gains from trade coincidence. Thus, it does not matter whether loss aversion is considered "a mistake" or "taken seriously".

The question whether some observed behavior is a mistake or not plays a crucial role in Chapter 2, which deals with the informational requirements of nudging. A nudge is a policy intervention that attempts to improve a person's choices by only reframing the decision problem. Thus, nudging is an attempt to help people avoid making mistakes, thereby raising the question of how to identify mistakes. The existing literature has failed to provide a solid theoretical foundation for nudging and typically dismisses the question of what constitutes a mistake. In Chapter 2, Nick Netzer and I attempt to fill this gap and propose a choice-theoretic foundation for nudging by appropriately modifying the standard revealed preference approach. To do so, we introduce a welfare criterion which is based on the agent's preference and allows us to identify mistakes by observing choices under different frames. In this context, a frame encompasses all the aspects of a choice problem which are not welfare relevant. Further, a behavioral model determines how the frame affects the agent's behavior given her welfare preference. With this general framework in hand, we characterize classes of behavioral models for which nudging is

possible or impossible. When nudging is possible we further derive results on the required quantity of information. Most of our results paint a rather negative picture for nudging in spite of several modeling choices stacking the odds in favor of it. Finally, we present an extended application of our results to a savings problem.

In Chapter 3 Igor Letina and I study the optimal design of dynamic research contests. Contests have a long history as mechanisms for inducing innovation. Although research is inherently a dynamic process, most of the literature considers static research contests only. The notable exception is the seminal work on dynamic research contests by Taylor (1995), whose framework we closely mirror. The key departure from his framework is to introduce progress prizes. Thus, instead of having only a prize which goes to the winner at the end of the contest, the principal who initiated the contest has to pay out a progress prize in each period as long as the contest continues. Moreover, we allow the principal to choose the deadline and end the contest before actually reaching the deadline. It turns out that this minor departure dramatically changes the results. First, we show that it is possible to implement what is called a global stopping rule using such a progress-prize contest, a finding in sharp contrast with the individual stopping rule which Taylor (1995) can implement. Second, capitalizing on this result we show that we can implement the first-best outcome using a progress-prize contest, which thus constitutes the optimal mechanism in this setting. Third, if the length of the contest is for some reason exogenously given as in Taylor (1995), we show that we can still implement the first best in case of a natural breakthrough innovation structure.

As has hopefully become clear, all three chapters deal with the design of economic institutions and with the question of how they in turn affect the behavior of the involved economic agents. Economic institutions are pervasive in our society and a good understanding of them is thus an important factor in any attempt to improve welfare. With this in mind, I invite you to continue reading and learn more about how to design economic institutions.

Part II

Research Papers

1 Bilateral Trade with Loss-Averse Agents[1]

1.1 Introduction

In many situations people evaluate an outcome relative to some reference point. For instance, if a buyer expects a trade to go through, her willingness to pay for the good may increase (Ericson and Fuster, 2011). Relatedly, whether a house owner is willing to sell her house at some price may depend on whether or not that price is higher than the original purchase price (Genesove and Mayer, 2001). Evidence suggests that the most relevant type of reference dependence in preferences is loss aversion (see DellaVigna, 2009, for a survey).[2] Kahneman and Tversky's (1979) prospect theory established the importance of loss aversion early on, and the literature on this phenomenon has grown substantially since. In particular, a large body of literature finds evidence of loss aversion in trade situations.[3] In spite of this empirical evidence the question of the effects of loss aversion on trade has not been addressed by the theoretical literature. In this paper, we aim to fill this gap and study the bilateral trade problem under the assumption that the agents are loss averse. More specifically, using the influential model of expectations-based loss aversion by Kőszegi and Rabin (2006, 2007b) (henceforth KR), we study how the presence of loss aversion affects the gains from trade which can be realized as well as the revenue an intermediary can make in a bilateral trade setting.

In the bilateral trade problem, a privately informed seller wants to sell one unit of an indivisible good to a privately informed buyer. In the classic framework of Myerson and Satterthwaite (1983) (henceforth MS) both agents have quasi-linear utility over ownership of the good and money. We augment the model by allowing for both agents to have reference-dependent preferences as modeled in KR. More precisely, an agent derives the standard *material utility* from ownership of the good and money, and, in addition, experiences *gain-loss utility* with respect to both, money and ownership of the good, separately. The reference point, relative to which agents evaluate an outcome, is formed endogenously as the rational expectations over the outcome.[4] We employ the choice-acclimating personal equilibrium (CPE) introduced in Kőszegi and Rabin (2007b) as our equilibrium concept. Thus, agents take an optimal action, taking into account that this action determines their reference point and the eventual outcome.

In this framework, we can distinguish between material gains from trade, corresponding to those in the absence of loss aversion as in MS, and total gains from trade, which include the gain-

[1] This paper should be cited as Benkert, J.-M. (2017), "Bilateral Trade with Loss-Averse Agents," Mimeo.

[2] There is a substantial empirical evidence of loss aversion, e.g., Fehr and Goette (2007), Post, van den Assem, Baltussen, and Thaler (2008), Crawford and Meng (2011) and Pope and Schweitzer (2011).

[3] See Ericson and Fuster (2014) for an excellent review on the role of loss aversion in explaining behavioral effects in exchange situations.

[4] Ericson and Fuster (2011), Abeler, Falk, Goette, and Huffman (2011), Crawford and Meng (2011), Gill and Prowse (2012), Karle, Kirchsteiger, and Peitz (2015), and Bartling, Brandes, and Schunk (2015) provide evidence for the assumption that the reference point is determined by expectations.

loss utility arising from trade in addition to the material gains. With this distinction in mind, we conduct our analysis of the effects of loss aversion on the gains from trade. We first examine the possibility of realizing all material gains from trade. The famous impossibility result in MS shows that in the absence of loss aversion not all material gains from trade can be realized given incentive compatibility, individual rationality and budget balance constraints. We show that while the impossibility result cannot be reversed in the presence of loss aversion, the severity of the problem can be mitigated. To do so, we show that the minimal subsidy needed to induce trade whenever the buyer values the good more than the seller can decrease in the degree of loss aversion of the buyer. This positive effect of buyer loss aversion on trade has been documented empirically and has been referred to as the *attachment effect* (Ericson and Fuster, 2011). In contrast, we show that seller loss aversion always increases the severity of the impossibility problem. This negative effect of seller loss aversion has also been documented empirically and is known as the *endowment effect* (Thaler, 1980). Formally, the two effects arise because loss aversion decreases the information rent of the buyer and increases the information rent of the seller. The impossibility result cannot be reversed, however, because incentive compatibility puts constraints on how loss averse the agents can be, which limits the strength of the attachment effect. As we show at the end of the paper in a robustness section, this limiting effect of incentive compatibility on the strength of the attachment and endowment effect extends to other models of reference-dependent utility than the one considered in the main analysis.

The robustness of the impossibility result in the present context is in stark contrast to other papers with non-standard preferences which show that the impossibility result can be reversed. In the case of intentions-based social preferences the reversal is driven by the fact that the incentive compatibility constraints can be turned slack by introducing an action which generates sufficiently strong feelings of kindness, thereby essentially eliminating any tension between ex post efficiency and the agents' incentives (Bierbrauer and Netzer, 2016). Similarly, as agents become more altruistic, their utility becomes more aligned with the expected gains from trade, reducing the tension between ex post efficiency and the agents' incentives (Kucuksenel, 2012). Thus, in contrast to the present framework, the channel alleviating the impossibility problem does not conflict with the incentive compatibility or the individual rationality constraints, meaning that a reversal is possible.

Having confirmed the impossibility result in the presence of loss aversion, we turn to the problem of designing optimal mechanisms. We in turn consider the problem of maximizing the designer's revenue, and the problem of maximizing the expected gains from trade, both total and material only, subject to some budget constraint. We show that in the presence of loss aversion any mechanism maximizing revenue or gains from trade features what we call *interim-deterministic transfers*, that is, the transfer of an agent is independent of the other agent's report and is thus deterministic given her own type. This reduces ex-post variations in payoffs, thereby making loss-averse agents better off. Turning to the optimal trade rule (for both, revenue and gains from trade), we impose the assumption that types are drawn from the uniform distribution to keep the model tractable. In spite of this assumption it is not possible to obtain the optimal trade rule using pointwise maximization because the agents' expected utilities endogenously depend on the mechanism through the reference point. In order to nevertheless derive the

optimal trade rule we make use of the reduced-form approach. Border (1991) characterized which interim allocation probabilities are implementable by some ex post allocation rule in the case of single-unit auctions. Che, Kim, and Mierendorff (2013) substantially generalized this result to multi-unit auctions, and also extended the reduced-form approach from auctions to the bilateral trade setting. Thus, instead of maximizing over the ex-post trade rule, we can maximize directly over the interim trade probabilities subject to some feasibility constraints, allowing us to explicitly derive the optimal trade rule. We show that the designer optimally induces less trade in the presence of loss aversion. Thus, the designer eliminates all ex-post variation in the agents' transfers, thereby fully insuring them against any losses in the money dimension, and partially insures them against losses in the trade dimension by reducing the trade probability. Full insurance in the trade dimension boils down to trade always or never taking place, which is generally not optimal. For sufficiently high stakes and degrees of loss aversion, however, the designer indeed provides the agents with full insurance by eliminating trade altogether. Intuitively, as the stakes become larger, it becomes too costly to induce loss-averse agents to take on any uncertainty. Interestingly, we find that the mechanism maximizing the material gains from trade coincides with the mechanism maximizing the total gains from trade. Thus, it does not matter whether the designer treats loss aversion as a "mistake" and only cares about the material gains from trade, or, alternatively, takes loss aversion "seriously" and includes gain-loss utility in the gains from trade.

Our final results concern the robustness of the optimal mechanisms and of the impossibility result in the presence of loss aversion for other specifications of the formation of the reference point. In our analysis we employ the concept of a choice acclimating personal equilibrium (CPE), which, as KR note, is similar to models of disappointment aversion such as those introduced by Bell (1985) and Loomes and Sugden (1986). The CPE specifies the reference point endogenously as the full distribution of a lottery, whereas the reference point corresponds to the certainty equivalent of the lottery in these models of disappointment-aversion. Masatlioglu and Raymond (2016) find that the intersection of preferences induced by the CPE and any of the listed disappointment aversion models is simply expected utility. Thus, although the models seem to be very similar on first glance, the induced preferences are generically different. Nevertheless, we show that the optimal mechanisms derived in this paper for CPE are also optimal for the models by Bell (1985) and Loomes and Sugden (1986) and that the impossibility result continues to hold, too. Further, we briefly explore the possibility of an exogenously given fixed reference point. We model this using the framework from Spiegler (2012) where the agents have an exogenously given reference point and feel losses in case of negative deviations, but feel no gains in the case of positive deviations. We show that the impossibility result persists for a large range of parameters, for instance, whenever the degree of loss aversion is symmetric across the agents.

There are numerous theoretical papers working under the assumption of loss-averse agents, three of which are particularly closely related to ours.[5] Eisenhuth (2013) considers the problem

[5] Less closely related, de Meza and Webb (2007) consider incentive design under loss aversion, Gill and Stone (2010) model a two-player rank-order tournament when agents are loss-averse, Rosato (2014) proposes expectations-based loss aversion as an explanation for the "afternoon effect" observed in sequential auctions, and Karle and Peitz (2014) investigate firm strategy in imperfect competition.

of a risk-neutral seller who wants to maximize revenue by selling a good to loss-averse buyers. Using the framework of KR, he finds that the optimal auction is an all-pay auction with reserve price when agents bracket narrowly. This result corresponds to our finding that transfers are interim deterministic in optimal mechanisms and, as one can show, extends beyond the auction and bilateral trade setting. Rosato (2017) considers a sequential bargaining model with a risk-neutral seller and a loss-averse buyer.[6] In the framework of KR and assuming wide bracketing, he shows that the buyer's loss aversion softens the rent-efficiency trade off for the seller. Just as in the present paper, this is driven by the attachment effect: the buyer is willing to accept lower offers to avoid the risk of a breakdown of the negotiations. In contrast to the present paper, Eisenhuth (2013) and Rosato (2017) do not feature loss-averse sellers, but only loss-averse buyers. Using the dynamic model of reference-dependent utility in Kőszegi and Rabin (2009), Duraj (2015) considers the impact of news utility in mechanism design models.[7] In his framework, in addition to being loss averse over consumption utility, agents are also loss averse over changes in beliefs about their current and future consumption. In the context of bilateral trade, he shows on the one hand that, when the realization of the outcome is delayed, the extra slack in the incentive compatibility constraints due to news utility is enough to reverse the impossibility result, contrasting the robustness result in the present paper. On the other hand, he shows that the optimality of deterministic transfers in revenue-maximizing mechanisms in the present paper extends to the setting with news utility and a delayed realization of the outcome. In the case without delay, which proves to be more tractable than the setting with delay as well as the setting in the present paper, he solves for the revenue maximizing mechanism.

The classical bilateral trade model has received a lot of attention in the literature. Arguably, the departure from the setting in MS most closely related to our paper, is to consider risk-averse agents. Early on, Chatterjee and Samuelson (1983) showed that when agents "become infinitely risk averse" all material gains from trade can be realized using a double-auction. More recently, Garratt and Pycia (2015) examine the bilateral trade problem relaxing the assumption that the agents have quasi-linear utility.[8] Allowing for risk aversion and wealth effects, they provide conditions for the possibility of realizing all gains of trade. The impossibility result can be reversed in this setting, because the presence of risk aversion and wealth effects gives rise to additional gains from trade, which then suffice to cover the agents' information rent. To put this result in perspective to our paper, we should note that the notions of efficiency being used to determine whether or not all gains from trade are realized differ. When considering material gains from trade only, we are effectively using the efficiency notion from MS's classical setting with quasi-linear utility and find robustness of the impossibility result. When we consider the problem of maximizing the total gains from trade we are closer to the efficiency notion used in

[6] See Shalev (2002) and Driesen, Perea, and Peters (2012) for other approaches incorporating loss aversion to bargaining.

[7] Both Duraj (2015) and Duraj's master thesis, from which said paper evolved, have been made available to us through personal communication. We thank Niccolò Lomys for making the connection. In the master thesis, Duraj also derives some results in the framework of the present paper. In particular, imposing stronger symmetry assumptions than here, he proves the robustness of the impossibility result and the optimality of deterministic transfers in revenue maximizing mechanisms.

[8] See also the references in Garratt and Pycia (2015) for more work on the bilateral trade problem in the classic framework with quasi-linear utility following MS. Moreover, see Wolitzky (2016) and Crawford (2016) for analyses of the bilateral trade problem with maxmin and level-k agents, respectively.

Garratt and Pycia (2015). However, in contrast to them, we do not establish whether efficient trade with respect to the total gains from trade can be achieved, but approach the problem as one of finding the trade mechanism which maximizes the total gains from trade from an ex-ante perspective. Further, on a more conceptual level, a model of bilateral trade with risk-averse agents is not suited to study the behavioral effects such as the endowment and attachment effect which have been documented empirically and, as noted above, are typically associated with loss aversion.

This paper is organized as follows. In Section 1.2 we introduce the model, solution concept and notation used throughout the paper. In Section 1.3 we study the effect of loss aversion on the gains of trade and information rents in order to address the impossibility result. Section 1.4 contains the derivation of the revenue and welfare maximizing mechanisms. In Section 1.5 we show that these optimal mechanisms display robustness to the exact specification of the reference point and Section 1.6 concludes. All proofs are relegated to the appendix.

1.2 Model

1.2.1 Utility, Social Choice Functions and Mechanisms

The set of agents is given by $I = \{S, B\}$ where S and B denote seller and buyer, respectively. It is commonly known that the type of agent $i \in I$ has distribution F_i with full support on the set $\Theta_i = [a_i, b_i] \subset \mathbb{R}_+$, and is private information. Let $\Theta = \Theta_S \times \Theta_B$ and assume that Θ_S and Θ_B have a non-trivial intersection. We interpret the type of an agent as her valuation of the good.[9] A social alternative is given by $\mathbf{x} = (y, t_S, t_B) \in X = \{0, 1\} \times \mathbb{R}^2$, where y indicates whether or not trade takes place and t_S and t_B denote the respective transfers of the seller and buyer.

Following KR, we allow for the agents to be loss-averse in the trade and in the money dimension. That is, the buyer derives the standard material utility from obtaining and paying for the good, and additionally, the buyer feels weighted gain-loss utility with respect to getting the good as well as weighted gain-loss utility with respect to paying for the good. Loss-aversion is captured by value functions in the sense of Kahneman and Tversky (1979) given by

$$\mu_i^k(x) = \begin{cases} x & \text{if } x \geq 0, \\ \lambda_i^k x & \text{else,} \end{cases}$$

for some $\lambda_i^k > 1$, which reflects the degree of loss aversion.[10] Thus, the riskless total utility is given by

$$u_S(\mathbf{x}, \mathbf{r}_S, \theta_S) = (1 - y)\theta_S + t_S + \eta_S^1 \mu_S^1 \left(r_S^1 \theta_S - y\theta_S \right) + \eta_S^2 \mu_S^2 (t_S - r_S^2) \tag{1.1}$$

$$u_B(\mathbf{x}, \mathbf{r}_B, \theta_B) = y\theta_B - t_B + \eta_B^1 \mu_B^1 (y\theta_B - r_B^1 \theta_B) + \eta_B^2 \mu_B^2 (r_B^2 - t_B) \tag{1.2}$$

[9] We could alternatively assume that the seller does not own the good but has to produce it. The seller's type would then represent her marginal cost of production. All the results that follow would go through in this case.

[10] We follow the literature by abstracting from diminishing sensitivity.

where $\eta_i^k \geq 0$ are the weights put on gain-loss utility. The parameters $\mathbf{r}_i = \{r_i^1, r_i^2\}$ are the so-called riskless reference levels. Following KR we will allow the reference point to be the agent's rational expectations and therefore a probability distribution over all riskless reference levels (see more below). We will refer to $(1-y)\theta_S + t_S$ and $y\theta_B - t_B$ as material utility and to the other terms as gain-loss utility in the trade and money dimension, respectively.
We adopt the following assumption from Herweg, Müller, and Weinschenk (2010):[11]

Assumption 1.1 (No Dominance of Gain-Loss Utility) $\Lambda_i = \eta_i^1(\lambda_i^1 - 1) \leq 1$, $i \in I$.

This assumption ensures that gain-loss utility does not dominate material utility and plays an important role for incentive compatibility. In particular, KR show that this condition ensures that agents will not choose stochastically dominated options. We will maintain this assumption throughout the paper and discuss the implications of relaxing it after deriving the impossibility result in Section 1.3. We follow KR by assuming that there is a separate gain-loss term for each of the two material utility dimensions, trade and money utility.[12]

A social choice function (SCF) $f : \Theta \to X$ assigns a collective choice $f(\theta_S, \theta_B) \in X$ to each possible profile of the agents' types $(\theta_S, \theta_B) \in \Theta$. In the present bilateral trade setting, a social choice function takes the form $f = (y^f, t_S^f, t_B^f)$. Let $\mathcal{F}$ denote the set of all SCFs and $\mathcal{Y}$ the set of all trade mechanisms, i.e., the set containing all y^f. A mechanism $\Gamma = (M_S, M_B, g)$ is a collection of message sets (M_S, M_B) and an outcome function $g : M_S \times M_B \to X$. We denote the direct mechanism by $\Gamma^d = (\Theta_S, \Theta_B, f)$. Since agents privately observe their types, they can condition their message on their type. Consequently, a pure strategy for agent i in a mechanism Γ is a function $s_i : \Theta_i \to M_i$. Note that $g(s_S(\theta_S), s_B(\theta_B)) \in X$. Let S_i denote the set of all pure strategies of agent i. Further, we denote the truthful strategy $s_i^t(\theta_i) = \theta_i$. Throughout, the operator $\mathbb{E}_{-i}$ denotes the expectation over the random variables $\tilde{\theta}_{-i}$ taking the value θ_i as given.

1.2.2 Equilibrium Concept and Revelation Principle

We use the concept of an (interim) choice-acclimating personal equilibrium (CPE) introduced in Kőszegi and Rabin (2007b).[13] The set of all riskless reference levels is given by the set of all social alternatives X. Essentially, the set X captures all the outcomes that could materialize at the end of the agents' interaction. In a mechanism Γ, agent i's action induces a distribution over the set of social alternatives X, conditional on the other agent playing s_{-i}. It is this endogenously generated distribution over X that forms the agent's reference point, or rather, reference

[11] This condition is commonly imposed, see for instance de Meza and Webb (2007), Eisenhuth and Ewers (2012), Eisenhuth (2013), Karle and Peitz (2014), and Rosato (2014).

[12] The assumption that the loss aversion parameters are commonly known may seem restrictive. However, we are essentially assuming that the functional form of the utility function is common knowledge, thereby following for instance Maskin and Riley (1984) who assume in their study of optimal auctions with risk-averse buyers that the buyers' parameter of risk-aversion is commonly known. We briefly discuss relaxing the assumption in the conclusion.

[13] KR also introduce the unacclimating personal equilibrium (UPE). In the UPE the agent "maximizes expected utility taking the reference point as given", whereas in the CPE the agent "maximizes expected utility given that it determines both the reference lottery and the outcome lottery". KR note that the CPE is more appropriate when the uncertainty is resolved after the agent's decision. We thus believe that the CPE is the more natural equilibrium concept in our context, as the report of an agent determines the uncertainty she feels about the outcome given her beliefs about the other agent's type.

distribution in a CPE. Effectively, when an agent evaluates an outcome, she is comparing it to all other possible social alternatives that could have materialized given the distribution induced over them. Moreover, when the agent takes an action in a CPE, she takes the action anticipating that it will not only determine the outcome of the mechanism, but also the distribution over the set X and, therefore, the reference point.

Moving to the interim stage and allowing the reference point to be the agent's rational expectations, we can define the interim expected utility of the seller with type θ_S, in the mechanism Γ, when playing action $m \in M_B$, given that the buyer plays strategy s_B as

$$\begin{aligned}
U_S(m,&s_B,\Gamma|\theta_S) = \\
&\int_{a_B}^{b_B} (1 - y^g(m, s_B(\theta_B)))\theta_S + t_S^g(m, s_B(\theta_B))\, dF_B(\theta_B) \\
&+ \int_{a_B}^{b_B}\int_{a_B}^{b_B} \eta_S^1\mu_S^1\left(y^g(m, s_B(\theta'_B))\theta_S - y^g(m, s_B(\theta_B))\theta_S\right)\, dF_B(\theta'_B)\, dF_B(\theta_B) \\
&+ \int_{a_B}^{b_B}\int_{a_B}^{b_B} \eta_S^2\mu_S^2\left(t^g(m, s_B(\theta_B)) - t^g(m, s_B(\theta'_B))\right)\, dF_B(\theta'_B)\, dF_B(\theta_B) \\
&= \theta_S \int_{a_B}^{b_B} (1 - y^g(m, s_B(\theta_B)))\, d\theta_B + \int_{a_B}^{b_B} t_S^g(m, s_B(\theta_B))\, dF_B(\theta_B) \\
&+ \theta_S\eta_S^1 \int_{a_B}^{b_B}\int_{a_B}^{b_B} \mu_S^1\left(y^g(m, s_B(\theta'_B)) - y^g(m, s_B(\theta_B))\right)\, dF_B(\theta'_B)\, dF_B(\theta_B) \\
&+ \eta_S^2 \int_{a_B}^{b_B}\int_{a_B}^{b_B} \mu_S^2\left(t_S^g(m, s_B(\theta_B)) - t_S^g(m, s_B(\theta'_B))\right)\, dF_B(\theta'_B)\, dF_B(\theta_B).
\end{aligned} \tag{1.3}$$

The expression in (1.3) may require some explanation. The first line corresponds to material utility, the second to gain-loss utility in the trade dimension and the third to gain-loss utility in the money dimension. The double integral has a clear intuition. To illustrate, consider the third line containing the money gain-loss utility. Fix any θ_B in the domain of integration of the outer integral and suppose this was the actual realization of the buyer's type. The seller would then receive a transfer of $t_S^g(m, s_B(\theta_B))$, which she would compare to the reference point. The reference point, or rather distribution, is induced endogenously and corresponds to the distribution of possible transfers. Thus, for every θ'_B in the domain of the inner integral we get a possible transfer $t_S^g(m, s_B(\theta'_B))$ given the buyer's strategy and the seller's message. The seller compares the actual transfer $t_S^g(m, s_B(\theta_B))$ with all these other possible transfers and the value function μ_S^2 weights these comparisons differently, depending on whether they result in a loss or a gain. The inner integral then aggregates the gains and loss weighted by the induced probability distribution. Next, integrate over all the values θ_B in the domain of the outer integral to get the familiar interim expected utility. In summary, the seller aggregates over each possible realization of transfers and for each of these possible realizations she compares the outcome with all other possible outcomes, aggregating gains and losses in each comparison.

Given our interpretation that the seller owns the good, her outside option is type-dependent and given by θ_S. To simplify notation later, we will consider the seller's net utility from trade, which, with some abuse of notation, allows us to compactly write $U_S(m, s_B, \Gamma|\theta_S) = -\theta_S\tilde{v}_S(m)+$

$\tilde{t}_S(m)$, where

$$\begin{aligned}
\tilde{v}_S(m) &= \int_{a_B}^{b_B} y^g(m, s_B(\theta_B))\, dF_B(\theta_B) \\
&\quad - \eta_S^1 \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_S^1 \left(y^g(m, s_B(\theta'_B)) - y^g(m, s_B(\theta_B))\right)\, dF_B(\theta'_B)\, dF_B(\theta_B), \\
\tilde{t}_S(m) &= \int_{a_B}^{b_B} t_S^g(m, s_B(\theta_B))\, dF_B(\theta_B) \\
&\quad + \eta_S^2 \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_S^2 \left(t_S^g(m, s_B(\theta_B)) - t_S^g(m, s_B(\theta'_B))\right)\, dF_B(\theta'_B)\, dF_B(\theta_B).
\end{aligned}$$

This compact notation highlights the fact that not only material utility, but also overall utility is linear in the type. Moreover, it will turn out to be useful to further define

$$\begin{aligned}
\bar{t}_S(m) &= \int_{a_B}^{b_B} t_S^g(m, s_B(\theta_B))\, dF_B(\theta_B), \\
w_S(m) &= \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_B^2 \left(t_S^g(m, s_B(\theta_B)) - t_S^g(s_S(\theta_S), s_B(\theta'_B))\right)\, dF_B(\theta'_B)\, dF_B(\theta_B),
\end{aligned}$$

allowing us to write $\tilde{t}_S(m) = \bar{t}_S(m) + \eta_S^2 w_S(m)$. Similarly, we can write the buyer's utility as $U_B(m, s_S, \Gamma|\theta_B) = \theta_B \tilde{v}_B(m) + \tilde{t}_B(m)$, defining the functions $\tilde{v}_B$ and $\tilde{t}_B$ analogously.

We can now define our equilibrium concept, which follows Eisenhuth (2013).

Definition 1.1 *A strategy profile $s^* = (s_S^*, s_B^*)$ is a CPE of the mechanism $\Gamma = (M_S, M_B, g)$ if $s_i^*(\theta_i) \in \arg\max_{m_i \in M_i} U_i(m_i, s_{-i}^*, \Gamma|\theta_i)$ for all $i \in I$ and $\theta_i \in \Theta_i$.*

Definition 1.2 *A mechanism Γ implements a SCF f if there is a CPE strategy profile $s = (s_S, s_B)$ such that $g(s_S(\theta_S), s_B(\theta_B)) = f(\theta_S, \theta_B)$ for all $(\theta_S, \theta_B) \in \Theta$.*

Definition 1.3 *A SCF f is CPE incentive compatible (CPEIC) if the truthful profile $s^t = (s_S^t, s_B^t)$ is a CPE strategy in the direct mechanism Γ^d.*

As a first result we note that the revelation principle for CPE holds in our setting.

Proposition 1.1 (Revelation Principle for CPE) *A social choice function f can be implemented in CPE by some mechanism Γ if and only if f is CPEIC.*

The standard proof of the revelation principle goes through in spite of the presence of an endogenous reference point. To see this, note that the reference point is determined as the rational expectations over outcomes. Starting from an arbitrary mechanism which induces some distribution of outcomes, the corresponding direct mechanism induces the same distribution of outcomes and therefore also the same reference point. Henceforth, we focus on direct mechanisms and no longer explicitly list the mechanism as an argument in the utility function.

1.2.3 Incentive Compatibility and Efficiency

In this section we characterize the set of all CPEIC social choice functions and introduce some familiar concepts, such as individual rationality and ex post budget balance. Further, we introduce our notion of an interim deterministic mechanism.

Proposition 1.2 *The SCF $f = (y^f, t_S^f, t_B^f)$ is CPEIC if and only if,*

(i) $\tilde{v}_S$ is non-increasing and $\tilde{v}_B$ is non-decreasing, and

(ii) we can write utility as

$$U_S(\theta_S, s_B^t|\theta_S) = U_S(b_S, s_B^t|b_S) + \int_{\theta_S}^{b_S} \tilde{v}_S(t)\, dt, \tag{1.4}$$

$$U_B(\theta_B, s_S^t|\theta_B) = U_B(a_B, s_S^t|a_B) + \int_{a_B}^{\theta_B} \tilde{v}_B(t)\, dt. \tag{1.5}$$

The proof is standard and therefore omitted.[14] Recall that the functions $\tilde{v}_B$ and $\tilde{v}_S$ contain terms of gain-loss utility. Thus, while the incentive-compatibility conditions in Proposition 1.2 take the same form as in the absence of loss aversion, it need not follow that the set of incentive-compatible SCF coincides. We say that a SCF is individually rational if for both agents $i \in I$

$$U_i(\theta_i, s_{-i}^t|\theta_i) \geq 0 \quad \forall \theta_i \in \Theta_i, \tag{IR}$$

and that it is ex post budget balanced if

$$t_S^f(\theta_S, \theta_B) = t_B^f(\theta_S, \theta_B), \quad \forall (\theta_S, \theta_B) \in \Theta. \tag{BB}$$

Setting the outside option in (IR) equal to zero is without loss of generality.[15] An agent could choose to walk away and not participate in the mechanism as soon as she learns her type. Doing so would rule out any possibility of trade and payment or receipt of any transfers. Therefore, the reference points of the agent would be equal to zero, as she anticipates that no trade or transfers can take place if she walks away. Consequently, there would be no feelings of gain or loss, as well as zero material utility when the agent walks away.

We say that a mechanism has interim-deterministic transfers, when, given her own type, an agent's transfer does not depend on almost all types of the other agent. Similarly, a trade rule is interim deterministic, when, given her own type, the trade rule coincides for almost all types of the other agent. A mechanism with interim-deterministic transfers and an interim-deterministic trade rule is called interim deterministic.

1.3 Information Rents and Material Gains From Trade

The impossibility theorem in MS is commonly interpreted in terms of the difference between the material gains from trade and the informations rents: trade between the buyer and the seller does not create enough gains to cover the information rents that need to be given to the agents.

[14] In contrast to Carbajal and Ely (2016), who consider price discrimination using a different model of loss aversion than the one here, the standard integral representation obtains in our setting. This is driven by the fact that, in contrast to Carbajal and Ely (2016), the report of an agent and not her type determines her reference point. For instance, a high buyer type does not expect to get the good with the probability corresponding to her true type when misreporting. Rather, she is aware that reporting a lower type changes the probability of getting the good and this is reflected in her reference point.

[15] Recall that we are considering net utility and have thus already taken care of the seller's type-dependent outside option.

As a consequence, it is not possible to realize all material gains from trade under incentive compatibility and individual rationality without subsidizing the agents. In the light of this, we consider the question of how loss aversion affects the designer's ability to realize all material gains from trade. As already noted, two behavioral effects have been empirically documented: the attachment and the endowment effect (Ericson and Fuster, 2011; Thaler, 1980). The former, which relates to the buyer, facilitates trade, while the latter, which relates to the seller, impedes trade. In what follows, we will see that these empirical effects have theoretical counterparts working in precisely those directions.

Lemma 1.1 *Loss aversion decreases the total gains from trade of a mechanism if and only if the mechanism is not interim deterministic.*

The proof of the lemma is straightforward. Loss-averse agents dislike ex-post variations in their payoffs, which reduces their interim utility. Only in the case of an interim-deterministic mechanism, ex-post variations in the transfers and the trade outcome are completely eliminated (from an interim perspective) and therefore loss aversion does not decrease the total gains from trade.

The effect of loss aversion on the information rents is more subtle and interesting. We will now illustrate this using a simple mechanism. Consider the materially efficient trade rule $y^{ME}(\theta_B, \theta_S) = 1$ for $\theta_B \geq \theta_S$ and $y^{ME}(\theta_B, \theta_S) = 0$ for $\theta_B < \theta_S$ with transfers given by

$$\begin{aligned} t_B(\theta_S, \theta_B) &= -\theta_B \tilde{v}_B(\theta_B), \\ t_S(\theta_S, \theta_B) &= \theta_S \tilde{v}_S(\theta_S). \end{aligned}$$

This mechanism is special in two ways. First, under complete information, this mechanism fully extracts all rents from the agents. Hence, the mechanism is individual rational, but, as we will see momentarily, it is not incentive compatible. Second, the transfers are interim deterministic. Hence, the agents do not feel any gains or losses in the money dimension.

We begin by considering the effects of loss aversion on the buyer. The expected utility of reporting type θ'_B when θ_B is the agent's true type (and conditional on the seller reporting her type truthfully) is given by[16]

$$\begin{aligned} U_B(\theta'_B, s^t_S | \theta_B) &= \theta_B \tilde{v}_B(\theta'_B) - \tilde{t}_B(\theta'_B) \\ &= \underbrace{(\theta_B - \theta'_B) F_S(\theta'_B)}_{\text{material utility}} + \underbrace{\Lambda_B(\theta'_B - \theta_B)(1 - F_S(\theta'_B)) F_S(\theta'_B)}_{\text{gain-loss utility}}. \end{aligned} \tag{1.6}$$

In the classic framework of MS without loss aversion (i.e., with $\Lambda_B = 0$), a buyer of type θ_B would have an incentive to imitate a lower type θ'_B. This effect is still present as we can see from equation (1.6). Note that for the material utility we have $(\theta_B - \theta'_B) F_S(\theta'_B) > 0$ for $\theta_B > \theta'_B$, making a downward deviation profitable for the buyer in the same way as it does in the absence of loss aversion. However, loss aversion adds a new, countervailing effect: there is an incentive to imitate a *higher* type. When looking at the gain-loss utility in equation (1.6), we indeed have $\Lambda_B(\theta'_B - \theta_B)(1 - F_S(\theta_B)) F_S(\theta_B) > 0$ for $\theta_B < \theta'_B$. The intuition is as follows. Loss-

[16] We omit the derivations as they mirror the steps in the proof of Proposition 1.3 in Appendix A.1.1.

averse agents dislike payoff uncertainty. Since overall utility and, in particular, gain-loss utility is linear in the type, a higher buyer type dislikes the uncertainty more than a lower type. Recall that the mechanism we are considering in this example is fully rent-extracting. This allows us to decompose the transfer in two parts, one extracting the material utility, and the other extracting the gain-loss utility. When a buyer of type θ_B truthfully reports her type this yields a gain-loss utility of $-\Lambda_B\theta_B(1-F_S(\theta_B)F_S(\theta_B)$ and the corresponding component in the transfer is given by $\Lambda_B\theta_B(1-F_S(\theta_B)F_S(\theta_B)$ so that the gain-loss (dis-)utility is fully extracted. A deviation to a higher buyer type yields a transfer with a gain-loss component intended to extract the gain-loss utility of type, who values gains and losses more strongly. Thus, imitating a higher type is profitable, as it yields a transfer which compensates the gain-loss (dis-)utility of a higher buyer type and, therefore, leaves the buyer with some rent. The assumption that gain-loss utility does not dominate material utility ($\Lambda_B \leq 1$) ensures that overall the buyer still has an incentive to imitate a lower type. However, in the presence of loss aversion this incentive is diminished. As a consequence, the buyer's information rent is smaller in the presence of loss aversion. This reduction in the incentives to imitate a lower type and, in conjunction with that, the decrease in the information rent is precisely the attachment effect. Formally, and more generally, we can observe the reduction in the information rent of the buyer due to the attachment effect in an incentive compatible mechanism using the integral representation of the utility (see Proposition 1.2).

Turning to the seller, we can write the expected utility of reporting type θ'_S when θ_S is her true type as

$$\begin{aligned} U_S(\theta'_S, s^t_B|\theta_S) &= -\theta_S\tilde{v}_S(\theta'_S) + \tilde{t}_S(\theta'_S) \\ &= \underbrace{(\theta'_S-\theta_S)(1-F_B(\theta'_S))}_{\text{material utility}} + \underbrace{\Lambda_S(\theta'_S-\theta_S)F_B(\theta'_S)(1-F_B(\theta'_S))}_{\text{gain-loss utility}}. \end{aligned} \tag{1.7}$$

In contrast to the case of the buyer, the analogous exercise as above reveals that the presence of loss aversion *amplifies* the seller's incentive to imitate a high type. This increase in the incentives to imitate a higher type and in the information rent captures precisely the endowment effect. We summarize these findings in the following lemma.

Lemma 1.2 *Loss aversion in the trade dimension decreases the buyer's and increases the seller's information rent, respectively.*

The overall effect of loss aversion on the sum of information rents is ambiguous and as a consequence it is a priori unclear whether the impossibility result persists. As we will see, although the severity of the impossibility problem can be mitigated by loss aversion, it cannot be reversed. The result follows in two steps. First, observe that Lemmas 1.1 and 1.2 imply that it is sufficient to show the impossibility in the case when neither the seller nor the buyer are loss averse in the money dimension, and, moreover, the seller is not loss averse in the trade dimension either. To see this, note that loss aversion in the money dimension does not affect the agents' information rents, but may decrease the total gains from trade. Thus, loss aversion in the money dimension makes the problem unambiguously harder. The above discussion of the endowment effect showed that the seller's information rent increases in the presence of loss aversion in the trade

dimension. In addition, loss aversion in the trade dimension decreases the gains from trade, since the materially efficient trade rule is not interim deterministic. Thus, any loss aversion on the seller's side makes the problem unambiguously harder. Hence, it suffices to consider the case when the seller is not loss-averse and the buyer is loss-averse in the trade dimension only. Put differently, only the attachment effect could potentially reverse the impossibility result. Making use of this insight, the second step is to proceed analogously to the proof in MS. That is, impose budget balance as well as incentive compatibility to obtain an expression for the sum of utilities of the "worst" buyer and seller types in the materially efficient mechanism and show that it is strictly negative. Indeed, we obtain

$$\begin{aligned} &U_B(a_B) + U_S(b_S) = \\ &- \int_{\max\{a_B,a_S\}}^{\min\{b_S,b_B\}} (1 - F_B(x))F_S(x)(1 - \Lambda_B(1 - F_S(x))) + \Lambda_B(1 - F_S(x))F_S(x)xf_B(x)\, dx \quad (1.8) \\ &< 0, \end{aligned}$$

which violates individual rationality for any $\Lambda_B \leq 1$. This proves our first main result (see Appendix A.1.1 for the details).

Proposition 1.3 *Given CPEIC, IR and BB, it is impossible to realize all material gains from trade for any degree of loss aversion in the money or good dimension.*

The minimal subsidy needed to induce materially efficient trade under CPEIC and IR (see equation (1.8)) can be interpreted as a measure of the severity of the impossibility problem and will generally depend on the degree of loss aversion and the distribution of the agents' types. Indeed, taking the derivative of the minimal subsidy in equation (1.8) with respect to Λ_B, we can see that the attachment effect mitigates the impossibility problem by dominating the diminishing effect of loss aversion on the gains from trade whenever

$$\int_{\max\{a_B,a_S\}}^{\min\{b_S,b_B\}} (1 - F_B(x))F_S(x)(1 - F_S(x)) - (1 - F_S(x))F_S(x)xf_B(x)\, dx \geq 0.$$

To get a feel for this condition, consider the families of distributions $F_S(x) = x^s$ and $F_B(x) = x^b$ on $[0, 1]$ for $b, s > 0$. Whenever $b > 2s^2 - 1$ the buyer's loss aversion makes the problem easier. In words, the likelier low seller types and high buyer types are, the less severe is the impossibility problem. This is in line with the intuition underlying the attachment effect. When low seller types are likely, a buyer puts a relatively high probability on trade taking place and thus has a strong attachment to the good (a high reference point). Hence, when low seller types and high buyer types are likely, on average the buyer will have a high attachment effect, thereby mitigating the impossibility problem. Note that in the absence of loss aversion, it is also true that the minimal subsidy is lower the likelier low seller types and high buyer types are. In the presence of the attachment effect, however, this is reinforced.

Another noteworthy point is that for the extreme types, i.e., types who lie outside the intersection of the intervals, loss aversion does not matter. This finding is very intuitive. To see this, observe that for these types trade is interim deterministic and hence there is no gain-loss utility as there is no room for ex-post variations in payoffs. Put differently, expectations-based

loss aversion only has bite when there is unresolved uncertainty, which is only the case for types lying strictly in the intersection of the type spaces.

The fact that the impossibility result is not reversed is linked to the assumption that $\Lambda_B \leq 1$, i.e., that gain-loss utility does not dominate. For instance, when types are drawn from $[0,1]$ with distributions $F_S(x) = x$ and $F_B(x) = x^{10}$ the subsidy in equation (1.8) turns into a surplus for $\Lambda_B \geq 13/3$. However, in this example $\Lambda_B \leq 1$ is a necessary condition for the materially efficient mechanism to be incentive compatible for the buyer. Hence, incentive compatibility puts limits on the feasible degree of loss aversion, and, as a consequence, on the strength of the attachment effect, meaning that the impossibility result cannot be reversed. However, as we will discuss next, $\Lambda_B \leq 1$ is in general only a sufficient condition for incentive compatibility and not always necessary.

The assumption that $\Lambda_i \leq 1$ is commonly imposed in the literature for conceptual as well as technical reasons. In particular, KR showed that the assumption ensures that agents do not choose stochastically dominated options. In the present context, it is easy to show that the assumption is a sufficient condition for the materially efficient trade rule to be incentive compatible in the presence of loss aversion. Moreover, whenever $F_S(a_B) = 0$ the assumption is not only sufficient, but also necessary. That is, whenever the smallest buyer type has a zero probability of trading, the materially efficient trading rule is CPEIC if and only if $\Lambda_B \leq 1$. In particular, this is true when the types of both agents are drawn from the same support. It turns out, however, that when $F_S(a_B) > 0$ the assumption is no longer necessary.[17] Indeed, when $F_S(a_B) < 1/2$ the necessary condition reads $\Lambda_B \leq 1/(1 - 2F_S(a_B))$ and when $F_S(a_B) \geq 1/2$ no restrictions need to be put on Λ_B. In the light of the above result the question thus arises whether the impossibility result persists when $F_S(a_B) > 0$ and the assumption is relaxed, as this would allow us to strengthen the attachment effect and possibly set the required subsidy in equation (1.8) equal to zero.

To this end, one can show that the impossibility result continues to hold for $\Lambda_B \leq 1/(1 - F_S(a_B))$. This condition ensures that the lowest buyer type a_B is in fact the "worst" buyer type.[18] For $\Lambda_B > 1/(1 - F_S(a_B))$, the worst buyer type is some intermediate type and the above approach to proving the impossibility result fails: if the lowest buyer type is no longer the worst type, satisfying individual rationality for the lowest buyer type does no longer guarantee satisfying individual rationality for all types. The observation that an intermediate type is the worst type is reminiscent of the related model of partnership dissolution (Cramton, Gibbons, and Klemperer, 1987; Fieseler, Kittsteiner, and Moldovanu, 2003). In this model, the good is initially not exclusively owned by one agent only, but by several agents. As a result, the worst type of an agent may be an intermediate type. However, in spite of this similarity, the approach taken in that model cannot be extended to the present context due to the endogeneity of the reference point. In sum, although counterexamples have proved elusive, a reversal of the impossibility for when $\Lambda_B > 1/(1 - F_S(a_B))$ cannot be ruled out. Note, however, that for sufficiently high degrees of loss aversion the total gains from trade disappear completely. Thus,

[17] In Herweg et al. (2010), who first introduced this assumption, the assumption plays a similar role as here. It provides a sufficient but not necessary condition to satisfy incentive compatibility of certain contracts.

[18] Rosato (2014) makes the assumption that gain-loss utility does not dominate precisely to ensure that the lowest type of an agent is the worst type.

even if the buyer's information rent can be reduced using the attachment effect, impossibility will obtain for sufficiently high degrees of loss aversion because it will eliminate all the total gains from trade.[19]

1.4 Optimal Mechanisms

1.4.1 Maximizing the Designer's Revenue

The preceding section has confirmed the impossibility result in a framework with loss-averse agents under the standard assumption that gain-loss utility does not dominate. In particular, a designer who wants to realize all material gains from trade while satisfying incentive compatibility and individual rationality cannot make a positive profit. A natural question is thus whether a materially *in*efficient trade mechanism satisfying incentive compatibility and individual rationality can lead to a positive profit for the designer. To answer this question we consider the design of revenue maximizing mechanisms in the presence of loss-averse agents. We will first consider the case of general distributions and prove that the designer insures the agents against ex-post variations in their payoffs. More specifically, we show that in the presence of loss aversion optimal transfers are interim deterministic. We then restrict attention to the case where both the seller and buyer types are distributed uniformly on $[a,b]$ with $b=a+1$. The preceding, more general analysis of the impossibility result suggests that the symmetry of the type spaces is not a too restrictive assumption, as loss aversion does not matter for the extreme types for whom trade is interim deterministic. We focus on the uniform distribution for tractability and because it allows us to derive the trade rule explicitly.

The revenue-maximizing designer's problem reads

$$\max_{(y^f,t_S^f,t_B^f)\in\mathcal{F}} \int_{a_B}^{b_B}\int_{a_S}^{b_S}\left(t_B^f(\theta_S,\theta_B)-t_S^f(\theta_S,\theta_B)\right)\,dF_S(\theta_S)\,dF_B(\theta_B),$$

subject to CPEIC and IR. (RM)

We begin by rewriting this problem into a more accessible form which will allow us to gain some intuition first. The complete derivations and proofs of this section are contained in Appendix A.1.2. The first step is to impose the envelope representation of the utility due to the CPEIC and the individual rationality constraint. The objective function then reads

$$\int_{a_B}^{b_B}\left(\eta_B^2 w_B(\theta_B)+\theta_B\tilde{v}_B(\theta_B)-\int_{a_B}^{\theta_B}\tilde{v}_B(t)\,dt\right)\,dF_B(\theta_B)$$
$$+\int_{a_S}^{b_S}\left(\eta_S^2 w_S(\theta_S)-\theta_S\tilde{v}_S(\theta_S)-\int_{\theta_S}^{b_B}\tilde{v}_S(t)\,dt\right)\,dF_S(\theta_S).$$

In the absence of loss aversion, the envelope representation of utility would allow us to maximize

[19] In the above we have only discussed the degree of loss aversion of the buyer. Analogous arguments regarding the necessity and sufficiency of $\Lambda_S \leq 1$ for incentive compatibility of the seller apply. However, as loss aversion on the side of the seller makes the impossibility problem only harder, relaxing the assumption that gain-loss utility does not dominate does not affect our result.

over the trade rule only instead of both the trade rule and transfers. With loss aversion in the money dimension, however, this is not the case. Indeed, recall that we defined

$$w_S(\theta_S) = \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_S^2 \left(t_S^f(\theta_S, \theta_B) - t_S^f(\theta_S, \theta'_B) \right) \, dF_B(\theta'_B) \, dF_B(\theta_B),$$

and thus the objective function still depends on transfers. This expression and its analog for the buyer collect all gain-loss utility with respect to money. Nevertheless, the problem can be reduced to only choosing the optimal trade rule, because in any optimal mechanism the transfers of the seller will not depend on the buyer's type, and vice versa. To see this, note that

$$\begin{aligned}
w_S(\theta_S) &= \int_{a_B}^{b_B} \int_{a_B}^{b_B} \mu_S^2 \left(t_S^f(\theta_S, \theta_B) - t_S^f(\theta_S, \theta'_B) \right) \, dF_B(\theta'_B) \, dF_B(\theta_B) \\
&= \int_{a_B}^{b_B} \int_{a_B}^{b_B} \left(t_S^f(\theta_S, \theta_B) - t_S^f(\theta_S, \theta'_B) \right) \mathbb{1}[t_S^f(\theta_S, \theta_B) > t_S^f(\theta_S, \theta'_B)] \, dF_B(\theta'_B) \, dF_B(\theta_B) \\
&+ \int_{a_B}^{b_B} \int_{a_B}^{b_B} \lambda_S^2 \left(t_S^f(\theta_S, \theta_B) - t_S^f(\theta_S, \theta'_B) \right) \mathbb{1}[t_S^f(\theta_S, \theta_B) < t_S^f(\theta_S, \theta'_B)] \, dF_B(\theta'_B) \, dF_B(\theta_B) \\
&= \int_{a_B}^{b_B} \int_{a_B}^{b_B} \left(t_S^f(\theta_S, \theta_B) - t_S^f(\theta_S, \theta'_B) \right) \mathbb{1}[t_S^f(\theta_S, \theta_B) > t_S^f(\theta_S, \theta'_B)] \, dF_B(\theta'_B) \, dF_B(\theta_B) \\
&- \lambda_S^2 \int_{a_B}^{b_B} \int_{a_B}^{b_B} \left(t_S^f(\theta_S, \theta'_B) - t_S^f(\theta_S, \theta_B) \right) \mathbb{1}[t_S^f(\theta_S, \theta'_B) > t_S^f(\theta_S, \theta_B)] \, dF_B(\theta'_B) \, dF_B(\theta_B) \\
&= (1 - \lambda_S^2) \int_{a_B}^{b_B} \int_{a_B}^{b_B} \left(t_S^f(\theta_S, \theta'_B) - t_S^f(\theta_S, \theta_B) \right) \mathbb{1}[t_S^f(\theta_S, \theta'_B) > t_S^f(\theta_S, \theta_B)] dF_B(\theta'_B) dF_B(\theta_B),
\end{aligned}$$

where $\mathbb{1}$ denotes the indicator function. The key step in the above derivation lies in the last equality. Comparing the two integrands on the third and second-to-last lines, we notice that they look the same but that θ_B and θ'_B are interchanged. To see the equality, change the order of integration in the integral on the second-to-last line and perform a change of variables for the resulting integral. This shows that the two integrals are actually the same and allows us to sum them. Thus, since $\lambda_S^2 > 1$ we find $w_S(\theta_S) \leq 0$. As the expression enters the designer's maximization problem positively, she optimally sets $w_S(\theta_S) = 0$. Note that a transfer achieves $w_S(\theta_S) = 0$ if and only if the transfer is independent of almost all buyer types. Thus, interim deterministic transfers are the only transfers that achieve $w_S(\theta_S) = 0$. The argument for the transfers of the buyer is analogous.

Proposition 1.4 *Any solution to the revenue maximization problem* (RM) *entails interim-deterministic transfers.*

Intuitively, loss-averse agents dislike ex-post variations in their payoffs. By making the transfers independent of the other agent's type, the designer completely insures the agents from any ex-post variation in the transfers. Thus, starting from any mechanism with non-interim-deterministic transfers, the designer can extract more surplus from the agents by choosing appropriate interim-deterministic transfers, effectively selling the agents insurance. Note that interim-deterministic transfers are also *a* solution in the absence of loss aversion. However, in the presence of loss aversion interim-deterministic transfers are the *only* solution.[20]

[20] Eisenhuth (2013) proved an analogous result for the case of auctions. In fact, one can show that Proposition 1.4 extends beyond the bilateral trade and auction setting to general social choice functions. Further, the result is reminiscent of the optimal mechanism found in Herweg et al. (2010), who augment a principal-agent setting with

For the remainder of this section we will assume that the seller and buyer types are distributed uniformly on $[a, b]$ with $b = a + 1$ and explicitly derive the optimal trade rule. The assumption allows us to rewrite the maximization problem to

$$\max_{y^f \in \mathcal{Y}} \int_a^b (2\theta_B - 1 - a) y_B(\theta_B) \left(1 + \Lambda_B \left[y_B(\theta_B) - 1\right]\right) \, d\theta_B$$
$$- \int_a^b (2\theta_S - a) y_S(\theta_S) \left(1 - \Lambda_S \left[y_S(\theta_S) - 1\right]\right) \, d\theta_S, \tag{RM'}$$
$$\text{subject to } y_B(\theta_B) \text{ being non-decreasing and } y_S(\theta_S) \text{ being non-increasing,}$$

where $y_B(\theta_B) = \int_a^b y^f(\theta_S, \theta_B) \, d\theta_S$ and $y_S(\theta_S) = \int_a^b y^f(\theta_S, \theta_B) \, d\theta_B$ denote the interim trade probabilities of the buyer and seller, respectively. Let us inspect the objective function in (RM') more closely. The first integral corresponds to the expected payment the designer receives from the buyer and the second integral to the expected payment the designer makes to the seller. Note that the seller integral is always positive. The buyer integral is positive whenever $(2\theta_B - 1 - a) \geq 0$. Clearly, any optimal mechanism will therefore only induce trade for buyer types $\theta_B \geq (1 + a)/2$. Given this, both integrals are increasing in the trade probabilities y_B and y_S, respectively. Thus, the designer faces the intuitive trade-off that inducing trade comes at cost in the form of the payment due to the seller and with a benefit in the form of the payment from the buyer. Further, the form of the objective function suggests that even in the presence of loss aversion the designer wants to induce trade between high buyer and low seller types in particular. Put differently, the designer wants to buy the good from a low-value seller and sell it to a high-value buyer, as this yields a large profit margin. However, as a consequence of expectations-based loss aversion it matters for an agent's utility whether trade takes place with only a few or many types of the other agent, as this affects her expectations, which in turn determine her expected gain-loss utility. Thus, there are in some sense externalities between the outcomes of different types. Indeed, because the agents' expected utilities endogenously depend on the mechanism through the reference point, pointwise maximization of the objective function is not possible. In order to nevertheless explicitly derive the optimal trade rule, we make use of the reduced-form approach developed first by Border (1991) and recently generalized by Che et al. (2013). In the case of single-unit auctions, Border (1991) characterized which interim allocation probabilities are implementable by some ex-post allocation rule. Che et al. (2013) generalize this to the case of multi-unit auctions when agents may face capacity constraints. In particular, the results in Che et al. (2013) extend to the bilateral trade setting, allowing us to revert to this reduced-form approach. The conditions derived in Che et al. (2013) allow us to maximize directly over the interim trade probabilities y_B and y_S instead of the ex-post trade rule y^f. Using the conditions that ensure that these trade probabilities can actually be implemented by some ex-post trade rule, we can eliminate the seller's trade probability from the problem and maximize over y_B only. This allows us to transform the problem into one which can be solved using standard techniques from calculus of variations.

---

moral hazard by assuming the agent is expectations-based loss-averse as in the present paper. They find that the principal optimally employs a binary payment scheme instead of a fully contingent contract in the presence of loss aversion. Hence, loss aversion drastically reduces the ex-post variation payments, too, but, in contrast to the present setting, does not eliminate it fully to preserve incentives.

Proposition 1.5 *The revenue-maximizing trade rule is given by*

$$y^{RM}(\theta_S, \theta_B) = \begin{cases} 1 & \text{if } \theta_S \leq \delta^{RM}(\theta_B), \\ 0 & \text{otherwise.} \end{cases}$$

where δ^{RM} *is non-decreasing in* θ_B *and non-increasing in the parameters* Λ_S, Λ_B *and* a.

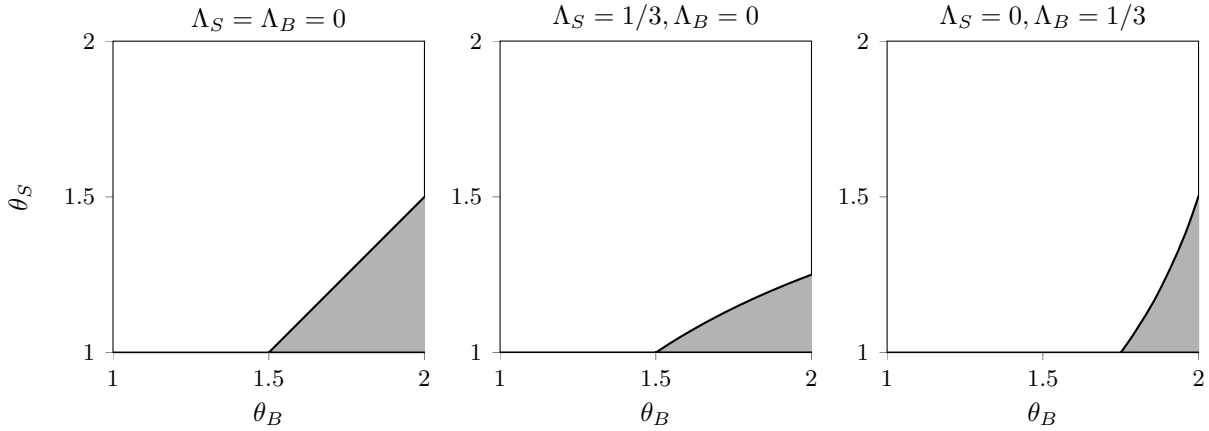


Figure 1.1: Illustration of the optimal trade rules for $a = 1$. The shaded area indicates for which pairs of types trade is taking place.

This result, for which the explicit expression can be found in the proof of the result in Appendix A.1.2, requires some discussion as it has several noteworthy features. First, in the absence of loss aversion in the trade dimension, i.e., for $\Lambda_S = \Lambda_B = 0$, we obtain the mechanism from MS in the framework without loss aversion given by $\delta^{RM}(\theta_B) = \theta_B - 1/2$. Second, the amount of trade taking place is monotonically decreasing in the degree of loss aversion and for sufficiently high degrees of loss aversion no trade takes place at all. Third, the trade-reducing effect of buyer loss aversion is stronger than the one of seller loss aversion.[21] This may come as a surprise in view of the endowment and attachment effect. In particular, when confirming the impossibility result under loss aversion, the endowment effect made the problem unambiguously harder, while the attachment effect had the potential to mitigate it, depending on the distribution of types. However, when types are distributed uniformly, the attachment effect does not mitigate the impossibility problem. Moreover, loss aversion affects the types of buyers and sellers the designer is most interested in differently. Indeed, the attachment and endowment effect are generally stronger for higher types, as these types value the gain-loss utility more strongly than low types. Moreover, as we already noted above, inducing trade increases the payment received from the buyer but it also increases the payment made to the seller. It is for this reason that the designer wants trade to take place in particular with high buyer types and low seller types. Hence, the effect of loss aversion is more pronounced for the buyer types than the seller types which are attractive from the revenue maximizing designer's point of view. Put differently, the adverse effect of loss aversion is increasing in the type of the agents. Since the designer cares

[21] For any value of a, as loss aversion increases, the buyer loss aversion will always lead to no trade taking place more quickly than seller loss aversion.

most about high buyer types and low seller types, buyer loss aversion has a stronger impact on the trade frequency than seller loss aversion.

Fourth, and perhaps most interestingly, the optimal mechanism depends on the type space. In the context of loss aversion, this suggest that the size of the stakes matters. In particular, for high stakes, i.e., high values of a, less trade takes place for any degree of loss aversion. This is in sharp contrast to the case without loss aversion, where the optimal mechanism is independent of the size of the stakes. Intuitively, the potential material gains from trade remain the same even when the stakes are high, because only the difference between valuation matters. However, as the stakes increase, the potential losses increase. Since the designer needs to compensate the agents for these losses with appropriate transfers to maintain individual rationality, the losses eventually eat up all the potential material gains. Hence, at some point the best the designer can do is to induce no trade at all. Contrary to conventional wisdom, the behavioral effects of loss aversion are not mitigated when the stakes are large. Rather, loss aversion has the biggest impact precisely when the stakes are large.

Finally, as already noted, by optimally making transfers interim deterministic, the designer provides the agents with insurance in the money dimension. Similarly, one can interpret the reduction in the trade dimension as partial insurance. Full insurance in this dimension would correspond to trade always or trade never taking place, which in general is not optimal. However, reducing the probability for trade lowers expectations and, as a consequence, there is less room for losses which benefits the agents.

1.4.2 Maximizing the Gains from Trade

In this section, we in turn consider the problem of maximizing total gains from trade and material gains from trade. In addition to CPEIC and IR, we impose a budget balance condition. Namely, we do not want the designer to inject money in the economy on average. This is in line with the preceding section, where we looked at ex-ante revenue maximization. We say that a mechanism is ex-ante budget balanced if

$$\int_{a_S}^{b_S}\int_{a_B}^{b_B}\left(t_S^f(\theta_S,\theta_B)-t_B^f(\theta_S,\theta_B)\right)\,dF_S(\theta_S)\,dF_B(\theta_B)=0. \tag{AB}$$

The problem of maximizing total gains from trade is given by

$$\max_{(y^f,t_B^f,t_S^f)\in\mathcal{F}}\int_{a_S}^{b_S}U_S(\theta_S,s_B^t|\theta_S)\,dF_S(\theta_S)+\int_{a_B}^{b_B}U_B(\theta_B,s_S^t|\theta_B)\,dF_B(\theta_B),$$
$$\text{subject to CPEIC, IR and AB.} \tag{TG}$$

To solve this problem, we proceed as we did in the preceding section. We also obtain the result that in any welfare maximizing mechanisms transfers will be interim deterministic.

Proposition 1.6 *Any solution to the problem of maximizing total gains from trade* (TG) *entails interim-deterministic transfers.*

The proof is analogous to the revenue maximization problem. In fact, just as in the case of revenue maximizing mechanisms, this result extends beyond the bilateral-trade setting and

applies to general social choice functions. To make further progress we again impose that types are uniformly distributed on $[a, b] = [a, a + 1]$. However, the presence of the budget constraint makes the problem less tractable, as we need to pin down the Lagrange multiplier. As a consequence, we need to impose symmetric degrees of loss aversion in the trade dimension, i.e., $\Lambda_B = \Lambda_S = \Lambda$. To derive the optimal trade rule we proceed as before for the revenue maximizing mechanism. That is, we make use of the reduced-form implementability conditions in Che et al. (2013) to derive the optimal interim trade probabilities. From there we recover an ex-post allocation rule which implements these probabilities and therefore is an optimal trade rule.

Proposition 1.7 *The trade rule maximizing total gains from trade is given by*

$$y^{TG}(\theta_S, \theta_B) = \begin{cases} 1 & \text{if } \theta_S \leq \delta^{TG}(\theta_B), \\ 0 & \text{otherwise}, \end{cases}$$

where δ^{TG} is non-decreasing in θ_B and non-increasing in the parameters Λ and a.

The optimal mechanism once more has some noteworthy features which qualitatively mirror those in the revenue maximizing mechanism. First, in the absence of loss aversion we get $\delta^{TG}(\theta_B) = \theta_B - 1/4$ which is the mechanism from MS in the framework without loss aversion. Second, loss aversion impedes trade. Third, the size of the stakes matter and as they increase the trade frequency diminishes until eventually no trade takes place at all. Finally, optimal transfers are interim deterministic. Thus, the designer optimally provides the agents with partial insurance in the trade dimension and with full insurance in the money dimension.

The trade rule y^{TG} with the corresponding transfers maximizes the total gain from trades. That is, it takes into account the gain-loss utility of the agents. Alternatively, the designer may be interested in maximizing only the material gains from trade, for instance, because she treats loss aversion as a behavioral mistake. This is captured by the problem

$$\max_{(y^f, t_B^f, t_S^f) \in \mathcal{F}} \int_{a_S}^{b_S} \left(-\theta_S y_S(\theta_S) + t_S(\theta_S)\right) \, dF_S(\theta_S) + \int_{a_B}^{b_B} \left(\theta_B y_B(\theta_B) - t_B(\theta_B)\right) \, dF_B(\theta_B),$$

$$\text{subject to CPEIC, IR and AB,} \tag{MG}$$

which maximizes only the material gains from trade, but the constraints respect the fact that agents themselves take gain-loss utility into account. The analysis of this problem is analogous to the problem of maximizing of total gains from trade. In fact, we obtain the following result.

Proposition 1.8 *The trade rule maximizing material gains from trade coincides with the trade rule maximizing total gains from trade.*

This result suggests a nice robustness property. Namely, it does not matter whether the designer considers gain-loss utility as part of the gains from trade or not. Further, the result implies that the trade frequency is reduced although the gain-loss utility does not enter the designer's objective function directly. However, as the constraints respect the fact that the agents are loss averse, gain-loss utility still enters the maximization problem through the incentive

compatibility and individual rationality constraints. The lack of a difference between the two mechanisms may still come as a surprise and is explained by the assumption that gain-loss utility does not dominate.

1.5 Alternative reference-point formation

The model by KR used in this paper has arguably become the workhorse model in the context of reference-dependent utility. A particularly appealing feature of the model is the endogenously determined reference point using the agent's rational expectations. As noted earlier (see footnote 4), a number of studies provide evidence for the assumption that a person's reference point is determined by her expectations. However, there are different ways one can model this. KR note that the equilibrium concepts in the models on disappointment aversion by Bell (1985) and Loomes and Sugden (1986) are closely related to the CPE. The CPE specifies the reference point as the full distribution of a lottery, whereas the reference point corresponds to the certainty equivalent of the lottery in these models of disappointment aversion. However, Masatlioglu and Raymond (2016) find that the intersection of preferences induced by the CPE and any of these disappointment-aversion models is only standard expected utility. Thus, although the models seem to be very similar, the induced preferences do generally not coincide. Nevertheless, the impossibility result in Section 1.3 remains valid and the optimal mechanisms derived in Section 1.4 coincide if we specify the reference point as the certainty equivalent of the lottery as in Bell (1985) and Loomes and Sugden (1986). Hence, the optimal mechanisms we derived earlier exhibits robustness to the specific formation of the reference-point.[22] To keep the analysis concise, we focus on the seller only. The arguments are essentially the same for the buyer. Under the alternative specification of the reference point the utility of the seller reads

$$\begin{aligned} U_S(\theta_S, s_B^t|\theta_S) = &\int_{a_B}^{b_B} \left(-y^f(\theta_S,\theta_B)\theta_S + t_S^f(\theta_S,\theta_B)\right)\, dF_B(\theta_B) \\ &+ \int_{a_B}^{b_B} \eta_S^1\mu_S^1 \left(\mathbb{E}_B[y^f(\theta_S,\tilde{\theta}_B)]\theta_S - y^f(\theta_S,\theta_B)\theta_S\right)\, dF_B(\theta_B) \\ &+ \int_{a_B}^{b_B} \eta_S^2\mu_S^2 \left(t_S^f(\theta_S,\theta_B) - \mathbb{E}_B[t_S^f(\theta_S,\tilde{\theta}_B)]\right)\, dF_B(\theta_B). \end{aligned}$$

Comparing this alternative expression to the expected utility we worked with (see equation (1.3)), we notice that the material utility on the first line remains unchanged, while the gain-loss utility in the second line takes a new form. Indeed, instead of comparing the induced outcome to every single potential outcome in the reference lottery, the agent now compares the outcome only to the certainty equivalent of the reference lottery, which enters the value function directly. Two observations about the alternative gain-loss utility yield the robustness result. Consider the money dimension first and recall that μ_S^2 is a concave function. Thus, by Jensen's

[22] Copic and Ponsatí (2008) have studied the bilateral trade problem in the context of robust mechanism design in the vein of Bergemann and Morris (2005). The robustness we have in mind here is closer to the behaviorally robust mechanisms in Bierbrauer and Netzer (2016).

inequality we get

$$\int_{a_B}^{b_B} \eta_S^2\mu_S^2\left(t_S^f(\theta_S,\theta_B)-\mathbb{E}_B[t_S^f(\theta_S,\tilde{\theta}_B)]\right)\,dF_B(\theta_B)$$
$$\leq \eta_S^2\mu_S^2\left(\int_{a_B}^{b_B}\left(t_S^f(\theta_S,\theta_B)-\mathbb{E}_B[t_S^f(\theta_S,\tilde{\theta}_B)]\right)\,dF_B(\theta_B)\right)=0,$$

as $\int_{a_B}^{b_B} t_S^f(\theta_S,\theta_B)\,dF_B(\theta_B)=\mathbb{E}_B[t_S^f(\theta_S,\tilde{\theta}_B)]$ by definition. Therefore, the result that $w_i(\theta_i)\leq 0$ carries through to this specification. Hence, irrespective of which of the two specifications of the reference point we use, interim-deterministic transfers are optimal.

Consider the trade dimension next and notice that $\mathbb{E}_B[y^f(\theta_S,\tilde{\theta}_B)]\in[0,1]$ while $y^f(\theta_S,\theta_B)\in\{0,1\}$. Thus, the binary nature of trade implies that an agent feels only either gains or losses in the trade dimension, irrespective of the reference lottery and outcome. We can thus rewrite

$$\begin{aligned}
&\int_{a_B}^{b_B}\eta_S^1\mu_S^1\left(\mathbb{E}_B[y^f(\theta_S,\tilde{\theta}_B)]\theta_S-y^f(\theta_S,\theta_B)\theta_S\right)\,dF_B(\theta_B)\\
&=\theta_S\eta_S^1\int_{a_B}^{b_B}\left(\lambda_S^1 y^f(\theta_S,\theta_B)\left(\mathbb{E}_B[y^f(\theta_S,\tilde{\theta}_B)]-1\right)+(1-y^f(\theta_S,\theta_B))\mathbb{E}_B[y^f(\theta_S,\tilde{\theta}_B)]\right)\,dF_B(\theta_B)\\
&=\theta_S\eta_S^1\int_{a_B}^{b_B}\int_{a_B}^{b_B}\left(\lambda_S^1 y^f(\theta_S,\theta_B)(y^f(\theta_S,\theta_B')-1)+(1-y^f(\theta_S,\theta_B))y^f(\theta_S,\theta_B')\right)\,dF_B(\theta_B')\,dF_B(\theta_B)\\
&=\theta_S\eta_S^1\int_{a_B}^{b_B}\int_{a_B}^{b_B}\mu_S^1(y^f(\theta_S,\theta_B')-y^f(\theta_S,\theta_B))\,dF_B(\theta_B')\,dF_B(\theta_B),
\end{aligned}$$

where the final line is the very expression of gain-loss utility in the trade dimension under the specification used throughout the paper. Thus, regarding gain-loss utility in the trade dimension the two different specifications of the reference point are equivalent.[23] Consequently, all our results continue to hold under the alternative specification of the reference point, as the two are equivalent conditional on interim deterministic transfers.

While the two formulations disagree on the precise way the reference-point is formed, they agree that it is the agents' expectations which determine the reference point endogenously. Alternatively, one could consider a model in which the reference point is exogenously given and not determined by the agent's expectations. We briefly explore this direction using the model of loss aversion used in Spiegler (2012) and reconsider the impossibility result in this framework. In the model by Spiegler (2012) agents have an exogenously given reference point r_i and feel losses in case of negative deviations, but they feel no gains in case of positive deviations. Thus, a buyer feels a loss of $\lambda_B r_B\theta_B$ when no trade happens, while the seller feels a loss of $\lambda_S(1-r_S)\theta_S$ when trade does happen. Similarly to the model by KR, loss aversion in the money dimension will only make the impossibility problem harder, as it decreases gains from trade without affecting information rents. We can write agents' expected utility as

$$U_B(\theta_B,r_B)=\theta_B y_B(\theta_B)-\bar{t}_B(\theta_B)-(1-y_B(\theta_B))\lambda_B r_B\theta_B$$

[23] This does not hinge on the piece-wise linearity of μ_i^1, but is solely due to the binary nature of trade.

and

$$U_S(\theta_S, r_S) = -\theta_S y_S(\theta_S) + \bar{t}_S(\theta_S) - y_S(\theta_S)\lambda_S(1 - r_S)\theta_S.$$

Collecting terms we observe that, as in the analysis in Section 1.3, seller loss aversion makes the problem unambiguously harder while the effect is ambiguous in case of the buyer. Hence, the endowment and attachment effect are once more at work. One can then follow essentially the same steps as we did for the proof of Proposition 1.3 to obtain that an incentive compatible, materially efficient, and budget balanced mechanism implies

$$\begin{aligned} &U_B(a_B) + U_S(b_S) = \\ &\int\int \left((1 + \lambda_B r_B)\left(\theta_B - \frac{1 - F_B(\theta_B)}{f_B(\theta_B)}\right) - (1 + \lambda_S(1 - r_S))\left(\theta_S + \frac{F_S(\theta_S)}{f_S(\theta_S)}\right)\right) y(\theta_S, \theta_B) dF_B(\theta_B) dF_S(\theta_S) \\ &- \lambda_B r_B \int \left(\theta_B - \frac{1 - F_B(\theta_B)}{f_B(\theta_B)}\right) dF_B(\theta_B). \end{aligned}$$

Thus, making use of the result in MS, one can see that a sufficient condition for the impossibility result to persist is given by $\lambda_B r_B \leq \lambda_S(1 - r_S)$. Whether the impossibility result extends in full generality, is not clear however.[24]

1.6 Conclusion

There are countless papers on mechanism design and vast evidence of the prevalence of loss aversion in people's behavior. Yet, as highlighted in a survey by Kőszegi (2014), work combining these two highly relevant fields is scarce. We contribute to this literature by investigating the bilateral trade problem with loss-averse agents. We first examine the possibility of realizing all material gains from trade and then derive mechanisms which maximize the designer's revenue as well as material and total gains from trade. We find that the presence of loss aversion generally impedes trade. Namely, a higher subsidy is required to induce materially efficient trade, the designer's revenue and the gains from trade which can be realized are reduced. The endowment and attachment effect, which are well-documented empirically, are apparent in our results and provide an intuitive explanation. The common theme in all three problems is that of insurance. In all optimal mechanisms interim-deterministic transfers are optimal, providing agents with full insurance in the money dimension. Additionally, less trade takes place in the presence of loss aversion, which can be interpreted as partial insurance in the trade dimension. Further, loss aversion affects the optimal mechanisms in a surprising and yet intuitive fashion. First, while both buyer and seller loss aversion reduce the optimal amount of trade, buyer loss aversion has a more pronounced impact, because loss aversion affects high types more strongly than low types, and the designer is particularly interested in high buyer types and low seller types. Second, the size of the stakes matter for the optimal mechanism: when the stakes are high, the designer optimally induces less trade, because the agents need to be compensated for risking large losses.

Interestingly and somewhat surprisingly, all of these findings display robustness to the exact

[24] Salant and Siegel (2016) study the efficient allocation of a divisible asset for different types of reallocation costs. For concave reallocation cost, the initial allocation can be interpreted as the reference point and deviations from the reference point lead to losses (but no gains) that are symmetric across agents. Thus, in this case their setting is very similar to the one here, but we allow for asymmetric losses across agents. In line with our findings, they show that ex-post efficiency may not be attained.

specification of the endogenous reference point. This is of practical relevance, as the designer of some economic institution may have evidence that individuals are loss averse, but be unsure about the precise formation process of the reference point. The robustness result suggests that lacking this information may not be too much of a problem, as long as loss-averse individuals are provided with insurance.

Throughout our analysis we have assumed that the degree of loss aversion is commonly known. If, instead, we assumed that these parameters are private information, a hard multi-dimensional mechanism design problem arises. Our analysis nevertheless provides some insights into this problem. We could relax the assumption that the loss-aversion parameters in the money dimension are commonly known and allow them to be distributed arbitrarily, as the designer optimally eliminates any ex-post variation in the transfers irrespective of the degree of loss aversion. We leave the question of private information regarding the degree of loss aversion in the trade dimension for further research.

Acknowledgments

I would like to thank Olivier Bochet, Juan Carlos Carbajal, Eddie Dekel, Jeff Ely, Samuel Häfner, Fabian Herweg, Heiko Karle, Botond Kőszegi, René Leal Vizcaíno, Igor Letina, Shou Liu, Daniel Martin, Konrad Mierendorff, Georg Nöldeke, Wojciech Olszewski, Anne-Katrin Roesler, Yuval Salant, Aleksei Smirnov, Ran Spiegler, Egor Starkov, Tom Wilkening, Peio Zuazo Garin, and seminar particpants in Zurich and at the ZWE 2014 for helpful comments. I am especially grateful to my supervisor Nick Netzer for his guidance as well as numerous comments and suggestions. I would like to thank the University of Basel and Northwestern University for their hospitality while some of this work was conducted and the UBS International Center of Economics in Society at the University of Zurich as well as the Swiss National Science Foundation (Doc.Mobility Grant P1ZHP1_161810) for financial support.

2 Informational Requirements of Nudging[1]

Joint with Nick Netzer

2.1 Introduction

A nudge (Thaler and Sunstein, 2008) is a regulatory intervention that is characterized by two properties. First, it is paternalistic in nature, because "it is selected with the goal of influencing the choices of affected parties in a way that will make those parties better off" (Thaler and Sunstein, 2003, p. 175). Second, it is not coercive but instead manipulates the framing of a decision problem, which makes it more easily acceptable than conventional paternalistic measures. Among the best-known examples already dicussed in Thaler and Sunstein (2003) is retirement saving in 401(k) savings plans, which can be encouraged tremendously by setting the default to automatic enrollment. Another example is the order in which food is presented in a cafeteria, which can be used to promote a more healthy diet.

The intriguing idea that choices can be improved by framing has made the concept of nudging also politically attractive. Governments of numerous countries have set up so-called "nudge units", which develop and implement nudge-based policies. The UK spearheaded this development in 2010 with the foundation of the Behavioral Insights Team.[2] More recently, US President Barack Obama issued an executive order establishing the Social and Behavioral Sciences Team.[3] The executive order encourages all government agencies to "carefully consider how the presentation and structure of [...] choices, including the order, number, and arrangement of options, can most effectively promote public welfare".

This paper addresses the problem of how to define and measure welfare. What does it mean that a frame improves choices? How can we be sure that it is in the employee's own best interest to save more or to eat more healthily? Due to behavioral inconsistencies caused by framing, the ordinary revealed preference approach is not suitable to answer these questions. Instead, the applied nudging literature often takes criteria such as increased savings or improved health for granted (see e.g. Goldin, 2015, for a discussion). Other authors have entirely dismissed the idea of nudging based on the welfare problem (see e.g. Grüne-Yanoff, 2012). We take a different, choice-theoretic approach. We investigate a framework where the welfare preference of an agent can be (partially) inferred from her choices under different frames, and the success of a nudge is evaluated on this basis. We thus attempt to develop a welfare-theoretic foundation for nudging in a revealed preference spirit, but appropriately modified. The twist is that, once we accept

[1] This paper should be cited as Benkert, J.-M. and N. Netzer (2016), "Informational Requirements of Nudging," *University of Zurich, Department of Economics, Working Paper No. 190.* A modified version of this paper has been submitted to the *Journal of Political Economy.*

[2] See http://www.behaviouralinsights.co.uk.

[3] See http://go.wh.gov/MKURtv.

that "in certain contexts, people are prone to error" (Sunstein, 2014, p. 4), we may be able to learn about these errors from choice data.[4]

Our formal framework is a variant of Rubinstein and Salant (2012), henceforth RS, who formulate a generalized approach for eliciting an agent's preferences from choice data. In this framework, which we formally introduce in Section 2.2, a regulator has a conjecture about the behavioral model d, which relates each pair of a *welfare preference* $\succeq$ and a *frame* f to a *behavioral preference* $d(\succeq, f)$. The interpretation is that an agent with welfare preference $\succeq$ acts as if maximizing $d(\succeq, f)$ if the situation is framed according to f. The welfare preference represents the normatively relevant well-being of the agent but is not observable. Behavioral preferences may be different from the welfare preference but are in principle observable in the usual revealed preference sense. RS investigate the problem of learning about the welfare preference from a data set that contains observations of behavior and, possibly, frames. We follow their approach in a first step, by verifying which welfare preferences could have generated a given data set. In a second step, we evaluate the frames based on the acquired information.

The framework's generality enables us to accommodate many different behavioral models. Among others, we will study well-known models such as choice from lists, default biases, satisficing, priming, and limited search. Our goal is not to take a stand on what the correct behavioral model is, or to argue in favor of any one of these models. Rather, the objective of our analysis is to understand the general properties of decision-making processes that make it possible or impossible to improve choices by framing.

A first contribution of our paper is to provide a choice-theoretic definition of a nudge. After identifying the welfare preferences that are consistent with a given data set and a behavioral model, in Section 2.3 we evaluate the frames on the basis of each of these preferences. Comparing frames pairwise, we say that a frame f is a weakly successful nudge over frame f' if the induced choices under f are at least as good as under f', irrespective of which of the consistent preferences is the actual welfare preference. This definition captures the above-mentioned idea that the regulator aims at improving the agent's choices by her own standards, i.e., the regulator tries to help the agent do what she really wants to do. It also shares with the literature (e.g. Masatlioglu et al., 2012) the cautious approach of requiring agreement among all possible welfare preferences, thereby ensuring that the regulator does not accidentally make the agent worse off.

Having formalized the concept of a successful nudge, we can formulate notions of global optimality. Ideally, we may be able to identify a frame that is a successful nudge over all the other frames. We show that the ability to identify such an optimal frame coincides with the ability to identify the welfare preference. An optimal frame is revealed by some sufficiently rich data set if and only if the welfare preference is fully revealed by some sufficiently rich data set. This does not mean that the welfare preference has to be fully elicited for successful nudging, as we will show by example, but it allows us to consider two polar cases: models in which the welfare preference can never be identified completely, and models in which the welfare

[4] Kőszegi and Rabin (2008b) first emphasized the possibility of recovering both welfare preferences and implementation mistakes from choice data, for a given behavioral model. Several contributions have studied this problem for specific models. Recent examples include Masatlioglu, Nakajima, and Ozbay (2012) for a model of limited attention, and Kőszegi and Szeidl (2013) for a model of focusing. Caplin and Martin (2012) provide conditions under which welfare preferences can be recovered from choice data in a setting where frames contain payoff-relevant information, such that framing effects are fully rational.

preference can be identified completely. There are interesting examples for either class of models, such as a satisficing model that has non-identifiable preferences and a limited search model that has identifiable preferences. We also ask how many models belong to each of the two classes and show that the share of models with identifiable preferences converges to 1, as the set of alternatives grows.

In Section 2.4 we investigate models with non-identifiable preferences more thoroughly. Finding an optimal frame is out of reach for these models, but we can still pursue the more modest goal of identifying frames which are dominated by others. Put differently, even though it is impossible to find a frame that improves upon all other frames, it may still be the case that some frames can be improved upon. Such dominated frames can indeed exist, as we show by example. However, if the behavioral model satisfies a property that we term the *frame cancellation property*, then all frames are always undominated, irrespective of the data set's richness. With the frame cancellation property, observation of choices never reveals the information required to improve these choices. Several important models have the frame cancellation property. A first example is the satisficing model in its different versions. A second example is the much-discussed case where the agent chooses the one alternative out of two that is marked as a default. We also present a decision-making procedure with limited sensitivity that nests all these (and more) behavioral models.

If, by contrast, the welfare preference can ultimately be learned, then questions of complexity arise. How many, and which, observations are necessary to determine the optimal frame? In Section 2.5 we define an *elicitation procedure* as a rule that specifies the order in which we impose different frames on the agent during an observation phase, contingent on the history of previous observations. This captures the idea that a data set may not be given randomly but can be collected deliberately with the purpose of finding an optimal nudge as quickly as possible. Holding fixed the unknown welfare preference of the agent, an elicitation procedure generates a sequence of expanding data sets. We define the complexity n of the nudging problem as the minimum over all elicitation procedures of the number of observations after which the optimal frame is guaranteed to be known. This number can sometimes be surprisingly small. For instance, we construct an optimal elicitation procedure for the limited search model and show that $n \leq 3$. We then establish a tight bound on n for arbitrary behavioral models. The bound, which is for instance reached by a behavioral model of priming, corresponds to the number of possible welfare preferences and thus grows more than exponentially in the number of alternatives. This implies that the informational requirements of nudging can in general become prohibitively large even with identifiable welfare preferences.

In Section 2.6 we allow for the possibility that the regulator has additional, non-choice-based prior information about the agent's welfare preference. We study such information in the form of restricted domains and of probabilistic beliefs over the set of preferences. For instance, the introduction of probabilistic beliefs allows us to generalize our notion of complexity in different ways. We investigate the expected running time of an elicitation procedure, and we relax our requirement of optimality and require that a frame is optimal only with a sufficiently large probability (or for a sufficiently large share of a population for which the agent is representative). As a consequence, nudging becomes easier, and sometimes substantially so.

In Section 2.7 we take the opposite direction and limit the regulator's exogenous information relative to the main model. We in turn relax the assumptions that the regulator has a unique conjecture about the correct behavioral model, thereby allowing for model uncertainty, and that the regulator can perfectly observe (and control) the frame under which the agent chooses. In the case of model uncertainty, for instance, the regulator needs to learn from choice data about both the welfare preference and the behavioral model. A fundamental new difficulty then arises when there are multiple model-preference pairs that are behaviorally equivalent but have different normative implications.

We present an extended application of our model to a savings problem in Section 2.8. We consider a two-period framework in which a risk-neutral agent derives utility from payments received in the present and in the future. Payments received in the future are discounted by some discount factor. In the spirit of our model, the discount factor is unobservable and represents the agent's welfare preference. There are two frames, one inducing a present and the other a future bias, respectively. We characterize when an optimal nudge exists and, if so, whether the agent should optimally be biased to the future or to the present. We then take this application to the data. To this end, we conducted an experiment on Amazon Mechanical Turk to elicit subjects' discount factors under a present-biased and a future-biased frame, respectively. Based on our model we can then estimate the subjects' underlying welfare discount factor and determine their nudgeability. We find a lot of heterogeneity in subjects' estimated welfare discount rates and for a large share of subjects our model predicts that an optimal nudge is given by inducing a present bias.

As noted before, the existing literature on nudging has focussed more on documenting the behavioral effects of framing, taking the welfare criterion for granted. We believe that our choice-theoretic approach adds a valuable new perspective. Several of our results imply strong informational limitations for a regulator who attempts to base the selection of nudges on a welfare-theoretic foundation. At the same time, our analysis reveals that seemingly minor differences between behavioral models – such as whether an agent's failure to optimize is due to a low aspiration level as in the satisficing model, or due to a restricted number of considered alternatives as in the limited seach model – can have profoundly different consequences for the ability to improve well-being by framing.

Goldin and Reck (2015) also study the problem of identifying welfare preferences when choices are distorted by frames, focussing mostly on binary choice problems with defaults. They estimate the preference shares among fully rational agents by the shares of agents who choose each alternative when it is not the default. The preference shares among the inconsistent agents are then deduced under identifying assumptions, for instance the assumption that they are identical to the rational agents after controlling for observable differences. It is then possible to identify the default that induces the best choice for a majority of the population. Informational requirements are not the only obstacle that a libertarian paternalist has to overcome. Spiegler (2015) emphasizes that equilibrium reactions by firms must be taken into account when assessing the consequences of a nudge-based policy. Even abstracting from informational problems, these reactions can wipe out the intended benefits of a policy. Finally, frames are often not chosen by a benevolent regulator but by profit-maximizing actors in markets, which also gives rise

to questions about welfare. Siegel and Salant (2015) study contracts when a seller is able to temporarily influence the buyers' willingness to pay by framing. They provide conditions under which optimal contracts make use of strategic framing, show how framing interacts with market regulation, and discuss the welfare implications.

2.2 Model and Examples

We begin by introducing the formal framework, which is a variant of RS, and we illustrate it with the help of two examples. Let X be a finite set of alternatives, with $m_X = |X|$. Denote by P the set of linear orders (reflexive, complete, transitive, antisymmetric) on X. A strict preference is a linear order $\succ \in P$. Let F be a finite set of frames, with $m_F = |F|$. By definition, frames capture all dimensions of the environment that can affect decisions but are not considered welfare-relevant.[5] The agent's behavior is summarized by a distortion function $d : P \times F \to P$, which assigns a distorted preference $d(\succ, f) \in P$ to each combination of $\succ \in P$ and $f \in F$. The interpretation is that an agent with true welfare preference $\succ$ acts as if maximizing the behavioral preference $d(\succ, f)$ if the choice situation is framed by f.[6] To fix ideas, we formally introduce two possible models.

Model 1 (Perfect-Recall Satisficing). This model is taken from RS. The agent is satisfied with any of the top k alternatives in her welfare preference, so $k \in \{2, \dots, m_X\}$ represents her aspiration level. The frame f describes the order in which the alternatives are presented to the agent. Whenever the agent chooses from some non-empty subset $S \subseteq X$ (e.g. the budget set), she considers the alternatives in S sequentially in their order as prescribed by $f \in F = P$. She chooses the first alternative that exceeds her aspiration level, i.e., she picks from S whichever satisfactory alternative is presented first. If S turns out not to contain any satisfactory alternative, the agent recalls all alternatives in S and chooses the welfare-optimal one. Choices between satisfactory alternatives will thus always be in line with the order of presentation, while all other choices are in line with the welfare preference. Hence we can obtain $d(\succ, f)$ from $\succ$ by rearranging the top k elements according to their order in f.[7]

Model 2 (Limited Search). This model formalizes a choice heuristic similar to one described in Masatlioglu et al. (2012). When the agent looks for a product online, all alternatives in X are displayed by a search engine, but only k of them on the first result page and $m_X - k$ of them on the second result page. The frame f here is the set of $k \in \{1, \dots, m_X - 1\}$ alternatives on the first page, such that F is the set of all size k subsets of X. The agent again chooses from non-empty subsets $S \subseteq X$ (e.g. not all displayed alternatives may be affordable to the agent or in

[5] For specific applications, the modeller has to judge which dimensions are welfare-relevant and which are not. For instance, it may be uncontroversial that an agent's well-being with some level of old age savings is independent of whether this level was chosen by default or by opt-in, but analogous statements would not be true if a default entails substantial switching costs, or if a "frame" actually provides novel information about the decision problem.

[6] This assumes that, given any frame, choices are consistent and can be represented by a preference. Salant and Rubinstein (2008) refer to (extended) choice functions with this property as "salient consideration functions" (p. 1291). The assumption rules out behavioral models in which choices violate standard axioms already when a frame is fixed. De Clippel and Rozen (2014) investigate the problem of learning from incomplete data sets without such an assumption.

[7] In contrast to RS, we explicitly treat the order of presentation as a variable frame. We also assume that the aspiration level k is fixed, which implies that the distortion function is single-valued.

stock with the retailer). Whenever the first result page contains at least one of the alternatives from S, then the agent does not even look at the second page but chooses from $S \cap f$ according to her welfare preference. Only if none of the elements of S is displayed on the first page, then the agent moves to the second page and chooses there according to her welfare preference. Choices between alternatives on the same page will thus always be in line with the welfare preference, but any available alternative on the first page is chosen over any alternative on the second page. Hence $d(\succeq, f)$ preserves $\succeq$ among all first and among all second page alternatives, but takes the first page to the top.[8]

The function d should be thought of as representing the regulator's conjecture about the relation between welfare, frames and choice. We consider the case of uncertainty about the behavioral model in Section 2.7, but for now we assume that the conjecture d is unique and given (keeping in mind that this assumption works in favor of nudging). Such a conjecture will typically rely on insights about the decision-making process and thus originates from non-choice data.[9] For instance, eye-tracking or the monitoring of browsing behaviors can provide the type of information necessary to substantiate a model like limited search (see the discussion in Masatlioglu et al., 2012), and methods from neuroscience may confirm decision-processes such as perfect-recall satisficing. As noted before, it is not our goal here to argue that a specific model is correct. Hence the only minor assumption that we impose on the behavioral model in general is that for each $\succeq \in P$ there exists an $f \in F$ such that $d(\succeq, f) = \succeq$. This rules out that some preferences are distorted by all possible frames and allows us to focus on the informational requirements of nudging, without having to deal with exogenously unavoidable distortions. The assumption does not imply the existence of a neutral frame that is non-distorting for all preferences.[10] In the satisficing model, all frames which present the k satisfying alternatives in their actual welfare order are non-distorting for that welfare preference. In the limited search model, the non-distorting frame places the k welfare-best alternatives on the first page.

Holding fixed a frame, the regulator now observes the agent's choices from sufficiently many different subsets $S \subseteq X$ to deduce her behavioral preference, in the usual revealed preference sense. Here the only difference to the usual approach is that the behavioral preference is not automatically equated with the welfare preference, and that the procedure generates potentially different revealed behavioral preferences when repeated for different frames. Formally, a data set is a subset $\Lambda \subseteq P \times F$, where $(\succeq', f') \in \Lambda$ means that the agent has been observed under frame f' and her choice behavior revealed the behavioral preference $\succeq'$. Further following RS, we say that $\succeq$ is consistent with data set Λ if for each $(\succeq', f') \in \Lambda$ it holds that $\succeq' = d(\succeq, f')$. In that case, $\succeq$ is a possible welfare preference because the data set could have been generated

[8] This model is also similar to the gradual accessibility model in Salant and Rubinstein (2008), but the eventual choice rule is different.

[9] Arguably, non-choice-based conjectures about the relation between choice and welfare always have to be invoked, even in standard welfare economics, see Kőszegi and Rabin (2007a, 2008a) and Rubinstein and Salant (2008). For an opposing perspective and a critical discussion of the ability to identify the decision process, see Bernheim (2009).

[10] Sometimes a neutral or "revelatory" frame (Goldin, 2015, p. 9) may indeed exist, for example when the default can be removed from a choice problem. The existence of such a frame makes the welfare elicitation problem and also the nudging problem straightforward. Often, however, this solution is not available, e.g. defaults are unavoidable for organ donations, alternatives must always be presented in some order or arrangement, and questions must be phrased in one way or another.

by an agent with that preference.[11] We illustrate the elicitation of the welfare preference, and also some first implications for nudging, using two examples.

Example 1. Consider an agent whose decision process is described by the perfect-recall satisficing model with aspiration level $k = 2$. The set of alternatives is given by $X = \{a, b, c, d\}$. The agent has the welfare preference $\succeq_1$ given by $c \succ_1 a \succ_1 b \succ_1 d$, so that alternatives c and a are satisfactory. Denote the frame which presents the alternatives in the alphabetical order by f. Thus, when choosing from some subset $S \subseteq X$, the agent will consider the alternatives in S in alphabetical order and choose the first which is satisfactory. Consequently, because a is presented before c, the agent will choose a whenever $a \in S$, even if also $c \in S$, in which case this is a mistake. She will choose c when $c \in S$ but $a \notin S$, and otherwise she will choose b over d by the perfect-recall assumption. Taken together, these choices look as if the agent was maximizing the preference $\succeq_2$ given by $a \succ_2 c \succ_2 b \succ_2 d$. Formally, we have $d(\succeq_1, f) = \succeq_2$. Suppose the behavioral preference $\succeq_2$ is observed in the standard revealed preference sense, by observing the agent's choices from different subsets $S \subseteq X$ but under the fixed frame of alphabetical presentation. Formally, the regulator obtains the data set $\Lambda = \{(\succeq_2, f)\}$. Given the perfect-recall satisficing conjecture, he can then conclude that the agent's welfare preference must be either $c \succ_1 a \succ_1 b \succ_1 d$ or $a \succ_2 c \succ_2 b \succ_2 d$; these two but no other welfare preferences generate the observed behavior under frame f. Formally, the set of preferences that are consistent with the data set is given by $\{\succeq_1, \succeq_2\}$. Therefore, with as little information as observing behavior under a single frame, the set of possible welfare preferences can be reduced from initially 24 to only 2.

We now illustrate some first implications for nudging, which here amounts to fixing an optimal order of presentation. Any order that presents a before c would be optimal if the agent's welfare preference was $a \succ_2 c \succ_2 b \succ_2 d$, but induces the above described decision mistake between a and c if the welfare preference is $c \succ_1 a \succ_1 b \succ_1 d$. The exact opposite is true for any order that presents c before a. Hence our knowledge is not yet enough to favor any one frame over another. Unfortunately, the problem cannot be solved by observing the agent under additional frames. The order of presentation fully determines choices among the alternatives a and c, so we can never learn about the welfare preference between the two. Since precisely this knowledge would be necessary to determine the optimal order, nudging here runs into irresolvable information problems.

Example 2. Consider an agent whose decision process is described by the limited search model, and $k = 2$ alternatives are presented on the first result page. As in the previous example, the set of alternatives is $X = \{a, b, c, d\}$ and the agent has the welfare preference $\succeq_1$ given by $c \succ_1 a \succ_1 b \succ_1 d$. Let $f = \{a, b\}$ denote the frame which puts the alternatives a and b on the first page. Thus, whenever the agent's choice set $S \subseteq X$ contains either a or b (or both), she will remain on the first page and make her choice there. Consequently, she chooses a whenever $a \in S$, even if also $c \in S$, because c is displayed only on the second page. This is again a mistake.

[11] Formally, this framework corresponds to the extension in RS where behavioral data sets contain information about frames. It simplifies their setup by assuming that any pair of a welfare preference and a frame generates a unique distorted behavioral preference. This is not overly restrictive, as the different contingencies that generate a multiplicity of distorted preferences can always be written as different frames. It is restrictive in the sense that observability and controllability of these frames might not always be given. See Section 2.7.2 for the respective generalization.

She will choose b when $b \in S$ but $a \notin S$, and otherwise she will choose c over d. Taken together, these choices look as if the agent was maximizing the preference $\succeq_3$ given by $a \succ_3 b \succ_3 c \succ_3 d$. Formally, we have $d(\succeq_1, f) = \succeq_3$. Suppose again that this behavioral preference is revealed, i.e., the regulator obtains the data set $\Lambda = \{(\succeq_3, f)\}$. Reversing the distortion process now unveils that the agent truly prefers a over b and c over d, which leaves the six possible welfare preferences marked in the first column of Table 2.1. The set of preferences consistent with the observed behavior is therefore given by $\{\succeq_1, \succeq_2, \succeq_3, \succeq_4, \succeq_5, \succeq_6\}$, meaning that the single observation reduces the set of possible welfare preferences from 24 to 6.

Here, an optimal nudge should place the two welfare-best alternatives on the first page, thus helping the agent avoid decision mistakes like the one between a and c under frame f above. Unfortunately, each of the four alternatives still belongs to the top two for at least one of the consistent welfare preferences, but none of them for all of the consistent welfare preferences. Hence no frame guarantees fewer mistakes than any other. In contrast to the satisficing example, however, gathering more information helps. Observing choices under frame $f' = \{a, d\}$ reveals the behavioral preference $\succeq_7$ given by $a \succ_7 d \succ_7 c \succ_7 b$, from which the welfare candidates marked in the second column of Table 2.1 can be deduced. Formally, adding this observation to the data set yields $\Lambda' = \{(\succeq_3, f), (\succeq_7, f')\}$, and the set of consistent welfare preferences shrinks to $\{\succeq_1, \succeq_2, \succeq_4, \succeq_5\}$. Note that these preferences all agree that a and c are the two best alternatives. Hence we know that $f'' = \{a, c\}$ is the optimal nudge. The actual welfare preference is still not known, so the example also shows that identifying a nudge is not the same problem as identifying the welfare preference.

	$f = \{a, b\}$: $a \succ_3 b \succ_3 c \succ_3 d$	$f' = \{a, d\}$: $a \succ_7 d \succ_7 c \succ_7 b$
$c \succ_1 a \succ_1 b \succ_1 d$	✓	✓
$a \succ_2 c \succ_2 b \succ_2 d$	✓	✓
$a \succ_3 b \succ_3 c \succ_3 d$	✓	
$a \succ_4 c \succ_4 d \succ_4 b$	✓	✓
$c \succ_5 a \succ_5 d \succ_5 b$	✓	✓
$c \succ_6 d \succ_6 a \succ_6 b$	✓	
$a \succ_7 d \succ_7 c \succ_7 b$		✓
$c \succ_8 b \succ_8 a \succ_8 d$		✓

Table 2.1: Reversing Limited Search

2.3 Nudgeability

2.3.1 Weakly Successful Nudge

In this section, we will provide a formal definition of a nudge. To capture the first step of preference elicitation due to RS in a concise way, let

$$\bar{\Lambda}(\succeq) = \{(d(\succeq, f), f) \mid f \in F\}$$

be the maximal data set that could be observed if the agent's welfare preference was $\succeq$, i.e., the data set that contains an observation for each possible frame. Then the set of all welfare preferences that are consistent with an arbitrary data set Λ can be written as

$$P(\Lambda) = \{\succeq \mid \Lambda \subseteq \bar{\Lambda}(\succeq)\}.$$

Without further mention, we consider only data sets Λ for which $P(\Lambda)$ is non-empty, i.e., for which there exists $\succeq$ such that $\Lambda \subseteq \bar{\Lambda}(\succeq)$. Otherwise, the behavioral model would be falsified by the data.[12] Observe that a frame f cannot appear more than once in such data sets. Observe also that $P(\varnothing) = P$ holds, and that $P(\Lambda) \subseteq P(\Lambda')$ whenever $\Lambda' \subseteq \Lambda$.

We are interested in evaluating the frames after having observed some data set Λ and having narrowed down the set of possible welfare preferences to $P(\Lambda)$. Since previously different frames may now have become behaviorally equivalent, let

$$[f]_\Lambda = \{f' \mid d(\succeq, f') = d(\succeq, f),\ \forall \succeq \in P(\Lambda)\}$$

be the equivalence class of frames for frame f, i.e., the elements of $[f]_\Lambda$ induce the same behavior as f for all of the remaining possible welfare preferences. We denote by $F(\Lambda) = \{[f]_\Lambda \mid f \in F\}$ the quotient set of all equivalence classes. Our central definition compares the elements of $F(\Lambda)$ pairwise from the perspective of the possible welfare preferences. For any $\succeq$ and any non-empty $S \subseteq X$, let $c(\succeq, S)$ be the element of S that would be chosen from S by an agent who maximizes $\succeq$.

Definition 2.1 *For any f, f' and Λ, $[f]_\Lambda$ is a weakly successful nudge over $[f']_\Lambda$, written*

$$[f]_\Lambda \; N(\Lambda) \; [f']_\Lambda,$$

if for each $\succeq \in P(\Lambda)$ it holds that $c(d(\succeq, f), S) \succeq c(d(\succeq, f'), S)$, for all non-empty $S \subseteq X$.

The statement $[f]_\Lambda N(\Lambda)[f']_\Lambda$ means that the agent's choice under frame f (and all equivalent ones) is at least as good as under f' (and all equivalent ones), no matter which of the remaining welfare preferences is the true one. The welfare preferences enter the definition not only for the evaluation of choices, but also because agents with different welfare preferences react differently to frames. The binary nudging relation $N(\Lambda)$ shares with other approaches in behavioral welfare economics the property of requiring agreement among multiple preferences (see, for instance, the multiself Pareto interpretation of the unambiguous choice relation by Bernheim and Rangel, 2009), but the multiplicity of preferences here simply reflects lack of information (as in Masatlioglu et al., 2012). Thus, adding observations to a data set can only make the partition $F(\Lambda)$ coarser and the nudging relation more complete, because it can only reduce the set of possible welfare preferences for which improved choices have to be guaranteed. In fact, the only way in which the data set Λ matters for the binary nudging relation is via the set $P(\Lambda)$.

The following Lemma 2.1 summarizes additional properties of $N(\Lambda)$ that will be useful. It relies on the sets of ordered pairs $B(\succeq, f) = d(\succeq, f) \setminus \succeq$ which record all binary comparisons

[12] RS derive conditions under which data sets do or do not falsify a model conjecture. A falsified model is of no use for the purpose of nudging and would have to be replaced by a conjecture for which $P(\Lambda)$ is non-empty.

that are reversed from $\succeq$ by f.[13] For instance, in the satisficing example in the preceding section, where the welfare preference was given by $c \succ_1 a \succ_1 b \succ_1 d$ and alphabetical order of presentation f resulted in the behavioral preference $a \succ_2 c \succ_2 b \succ_2 d$, we would obtain $B(\succeq_1, f) = \succeq_2 \setminus \succeq_1 = \{(a, c)\}$. For the limited search example where frame $f = \{a, b\}$ distorted the same welfare preference to $a \succ_3 b \succ_3 c \succ_3 d$, we would obtain $B(\succeq_1, f) = \succeq_3 \setminus \succeq_1 = \{(a, c), (b, c)\}$.

Lemma 2.1 *(i)* $[f]_\Lambda N(\Lambda)[f']_\Lambda$ *if and only if* $B(\succeq, f) \subseteq B(\succeq, f')$ *for each* $\succeq \in P(\Lambda)$.
(ii) $N(\Lambda)$ *is a partial order (reflexive, transitive, antisymmetric) on* $F(\Lambda)$.

The proof of the lemma (and all further results) can be found in Appendix B.1. Since $B(\succeq, f)$ describes all the mistakes in binary choice that frame f causes for welfare preference $\succeq$, statement (i) of the lemma formalizes the intuition that a successful nudge is a frame that guarantees fewer mistakes. Statement (ii) implies that the binary relation is sufficiently well-behaved to consider different notions of optimality.

2.3.2 Optimal Nudge

A benevolent regulator would ideally like to choose a frame that is a weakly successful nudge over all other frames and thus guarantees the best possible choices. We call such a frame an *optimal nudge*. Given a data set Λ, let

$$G(\Lambda) = \{f \mid [f]_\Lambda N(\Lambda)[f']_\Lambda, \forall f' \in F\}$$

be the set of frames which have been identified as optimal. Formally, $G(\Lambda)$ coincides with the greatest element of the partially ordered set $F(\Lambda)$, and it might be empty due to incompleteness of the binary nudging relation. Since the nudging relation becomes more complete as we collect additional observations, it follows that optimal nudges are more likely to exist for larger data sets. Therefore, the following result provides a necessary and sufficient condition for the existence of an optimal nudge for maximal data sets. The result is relatively straightforward but important, as it will allow us to classify behavioral models according to whether the search for an optimal nudge is promising or hopeless.

Definition 2.2 *Preference* $\succeq$ *is identifiable if for each* $\succeq' \in P$ *with* $\succeq' \neq \succeq$, *there exists* $f \in F$ *such that* $d(\succeq, f) \neq d(\succeq', f)$.

Proposition 2.1 $G(\bar{\Lambda}(\succeq))$ *is non-empty if and only if* $\succeq$ *is identifiable.*

The if-statement is immediate: an identifiable welfare preference is known once the maximal data set has been collected, and all the non-distorting frames are optimal with that knowledge. It is worth emphasizing again, however, that the result does not imply that the welfare preference actually has to be learned perfectly for successful nudging. It only tells us that, if $\succeq$ is the true and identifiable welfare preference, then for some sufficiently large data set Λ we will be able to identify an optimal nudge. The set $P(\Lambda)$ might still contain more than one element at that

[13] Even though we often represent preferences as rankings like $c \succ a \succ b \succ d$, we remind ourselves that technically both $d(\succeq, f)$ and $\succeq$ are subsets of the set of ordered pairs $X \times X$.

point. The only-if-statement tells us that there is no hope to ever identify an optimal nudge if the welfare preference cannot be identified, i.e., if there exists another welfare preference $\succeq'$ that is behaviorally equivalent to $\succeq$ under all frames. In this case we say that $\succeq$ and $\succeq'$ are *indistinguishable*. A frame could then only be optimal if it does not distort any of the two, but this is impossible as such a frame would generate different observations for $\succeq$ and $\succeq'$ and hence would empirically discriminate between them.

In the following, we will make use of the result in Proposition 2.1 and consider the two polar classes of behavioral models where all welfare preferences are identifiable or non-identifiable, respectively. Before turning to a detailed analysis of these two classes, we address the question of how plausible each of them is. Our prime example for non-identifiable preferences is the perfect-recall satisficing model. Any two welfare preferences that are identical except that they rank the same best k alternatives differently are mapped into the same distorted preference by any frame, and hence are indistinguishable. Our prime example for identifiable preferences is the limited search model (for $m_X \geq 3$). There, we learn the welfare preference among all alternatives on the same page, and thus we can identify the complete welfare preference by observing behavior under sufficiently many different frames. The decision processes formalized by these two models are both plausible, implying that both classes are important. Another way of looking at the question of plausibility is to ask how many models belong to each of the classes. We can provide an answer to this question for the limiting case as the number of alternatives grows large.[14] With m_X alternatives, there are $m_P(m_X) = m_X!$ strict preferences. The number of models also depends on how many frames $m_F(m_X)$ we allow, as a function of the number of alternatives. This number should typically be increasing in m_X, but for the following result we only need to assume that $m_F(m_X) \geq 4$ for sufficiently large values of m_X.[15]

Proposition 2.2 *The share of models with identifiable preferences goes to 1 as* $m_X \to \infty$.

The proof exploits the fact that the number of models with identifiable preferences is given by the number of different ways to assign distinct maximal data sets to the welfare preferences, satisfying the requirement that there must exist a non-distorting frame for each preference. It is difficult to determine this number exactly, but we find a lower bound that is tractable and suffices to show that the share of models with identifiable preferences converges to 1 as the number of alternatives grows. If one accepts that the genericity notion formalized by Proposition 2.2 captures model plausibility in a meaningful way, the result is good news for the nudging project. If the number of alternatives is large, an optimal nudge can generically be identified. However, we need to add that the complexity of finding this optimal nudge may become prohibitive if m_X is large, a problem to which we will return in Section 2.5.

[14] This approach of quantifying plausibility is similar to Kalai, Rubinstein, and Spiegler (2002), who are interested in the number of preferences that are necessary to rationalize an arbitrary choice function. They show that the share of choice functions which can be rationalized by less than the maximal conceivable number of preferences goes to 0 as the number of alternatives grows large.

[15] If we restricted attention to models where frames are orders of presentation, we would already obtain $m_F(m_X) = m_X!$. In general, the number of frames can be arbitrarily large. However, there can never be more than $m_F(m_X) = m_X!^{m_X!}$ different non-equivalent frames, the number of mappings from P to P.

2.4 Non-Identifiable Preferences

We now investigate behavioral models with non-identifiable preferences more thoroughly. From Proposition 2.1 we know that an optimal nudge cannot be found for these models. However, our previous notion of optimality was strong, requiring an optimal frame to outperform all other frames. Even if such a frame does not exist, we might still be able to exclude some frames that are dominated by others. We now weaken optimality to the requirement that a reasonable frame should not be dominated. Let

$$M(\Lambda) = \{f \mid [f']_\Lambda N(\Lambda)[f]_\Lambda \text{ only if } f' \in [f]_\Lambda\}$$

be the (always non-empty) set of frames which are undominated, based on our knowledge from the data set Λ. Formally, $M(\Lambda)$ is the union of all elements that are maximal in the partially ordered set $F(\Lambda)$. To provide an analogy, we can think of $M(\Lambda)$ as the set of Pareto efficient policies, because moving away from any $f \in M(\Lambda)$ makes the agent better off with respect to some $\succeq \in P(\Lambda)$ only at the cost of making her worse off with respect to some other $\succeq' \in P(\Lambda)$. By the same token, a frame which is not in $M(\Lambda)$ can be safely excluded, as there exists a nudge that guarantees an improvement over it.

Dominated frames can exist already ex ante with no knowledge of the agent's welfare preference. For instance, certain informational arrangements could be interpreted as being dominant over others, because they objectively clarify the available information and improve the decision quality (e.g. Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin, 2003). In the following example we show that ex ante undominated frames can become dominated for richer knowledge, too.

Example 3. Assume that $X = \{a, b, c, d\}$ and consider the distortion function for the four preferences and three frames depicted in Figure 1.[16] The two preferences $\succeq_1$ and $\succeq_2$ are indistinguishable, as each frame maps them into the same distorted preference, and the same holds for $\succeq_3$ and $\succeq_4$. Note also that none of the frames is dominated before any data has been collected, $M(\varnothing) = \{f_1, f_2, f_3\}$, because each one is the unique non-distorting frame for one possible welfare preference. Now suppose we observe $\Lambda = \{(\succeq_2, f_2)\}$, so that $P(\Lambda) = \{\succeq_1, \succeq_2\}$. It follows immediately that none of the potentially non-distorting frames f_1 and f_2 is dominated. The frame f_3, however, is now dominated by f_1. If the welfare preference is $\succeq_2$, then f_1 induces a mistake between a and b, but so does f_3, which induces an additional mistake between c and d. Hence we obtain $M(\Lambda) = \{f_1, f_2\}$. We have learned enough to identify a nudge over f_3, but no additional observation will ever allow us to compare f_1 and f_2.

The sometimes dominated frame f_3 in Example 3 has a particular property. It maps the indistinguishable set of preferences $\{\succeq_1, \succeq_2\}$ outside of itself. This is the reason why the example violates the following property.

[16] The example focusses on only four welfare preferences, but it can be expanded to encompass the set of all possible preferences. We can also add additional frames without changing its insight.

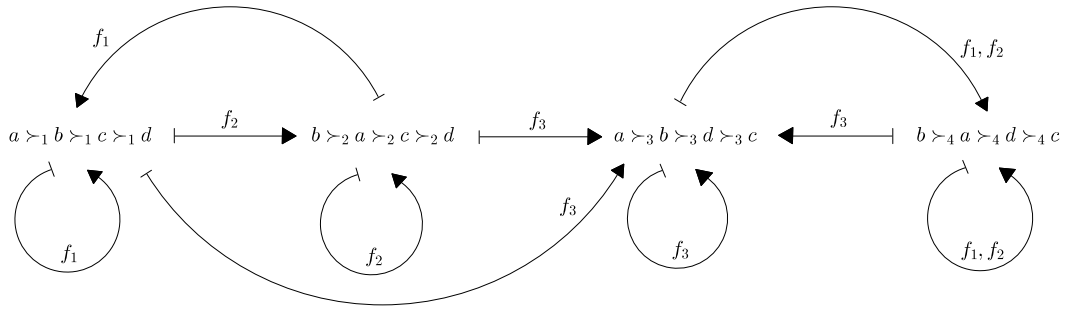


Figure 2.1: Dominated Frame f_3

Definition 2.3 *A distortion function d has the frame-cancellation property if*

$$d(d(\succeq, f_1), f_2) = d(\succeq, f_2)$$

holds for all $\succeq \in P$ and all $f_1, f_2 \in F$.

With the frame-cancellation property, the impact of any frame f_1 disappears once a new frame f_2 is applied. Starting from any welfare preference $\succeq$, the preference $d(\succeq, f)$ obtained by applying any frame $f \in F$ is then always observationally equivalent to $\succeq$, and thus is itself an indistinguishable welfare preference. Hence, for any given frame, all maximal indistinguishable sets of preferences are closed under the distortion function, in contrast to Example 3.

A variety of interesting behavioral models has the frame-cancellation property. One extreme example, where frames never have an effect on behavior and $d(\succeq, f) = \succeq$ always holds, is the rational choice model.[17] The opposite extreme case of frame-cancellation arises when $d(\succeq, f)$ is independent of $\succeq$, so that frames override the preference entirely. This is true, for instance, when there are only two alternatives and the agent always chooses the one that is marked as the default. The perfect-recall satisficing model has the frame-cancellation property, too, even though the welfare preference retains a substantial impact on behavior. In this model, the effect of the order of presentation is to overwrite the welfare preference among the top k alternatives, which leaves no trace of previous frames when done successively. We can also establish a connection to the analysis of choice from lists by Rubinstein and Salant (2006). They allow for the possibility that agents choose from lists instead of sets, i.e., the choice from a given set of alternatives can be different when the alternatives are listed differently. Their results imply that we can capture choice from list behavior in reduced form of a distortion function whenever the axiom of "partition independence" is satisfied by the agent's choices for all possible welfare preferences.[18] An example in which this holds is satisficing without recall. In contrast to the

[17] Note that all welfare preferences are identifiable in the rational choice model, which constitutes of course the basis for the standard revealed preference approach. The rational choice model is indeed the only model which has both identifiable preferences and the frame-cancellation property. To see why, suppose d is not fully rational, i.e., there exist $\succeq'$ and f' such that $d(\succeq', f') = \succeq'' \neq \succeq'$. If d has the frame-cancellation property we then obtain $d(\succeq'', f) = d(d(\succeq', f'), f) = d(\succeq', f)$ for all $f \in F$, hence $\succeq'$ and $\succeq''$ are indistinguishable and not identifiable.

[18] Partition independence requires that the choice from two concatenated sublists is the same as the choice from the list that concatenates the two elements chosen from the sublists (Rubinstein and Salant, 2006, p. 7). Such

perfect-recall version, the agent here chooses the last alternative on a list when no alternative on the list exceeds her aspiration level. Formally, $d(\succeq, f)$ is obtained from $\succeq$ by rearranging the top k elements in the order of f and the bottom $m_X - k$ elements in the opposite order of f (see RS). It is easy to verify that this model also has the frame-cancellation property. The following general class of decision processes nests all these models with the frame-cancellation property.

Model 3 (Limited Sensitivity). The agent displays limited sensitivity in the sense that she can sometimes not tell whether an alternative is actually better than another. Degree and allocation of sensitivity are described by a vector $(k_1, k_2, \ldots, k_s)$ of positive integers with $\sum_{i=1}^{s} k_i = m_X$. A welfare preference $\succeq$ induces a partition of X, where block X_1 contains the k_1 welfare-best alternatives, X_2 contains the k_2 next best alternatives, and so on. The agent can distinguish alternatives across but not within blocks. When choosing from $S \subseteq X$, she therefore only identifies the smallest i for which $S \cap X_i$ is non-empty, and the frame then fully determines the choice from this set. Thus $d(\succeq, f)$ is obtained from $\succeq$ by rearranging the alternatives within each block of the partition in a way that does not depend on their actual welfare ranking. Formally, let $P_{\succeq}$ be the set of welfare preferences that induce the same partition of X as $\succeq$, for any $\succeq \in P$. Then $d(\succeq', f) = d(\succeq'', f) \in P_{\succeq}$ must hold whenever $\succeq', \succeq'' \in P_{\succeq}$, for all $f \in F$. Any such function satisfies the frame-cancellation property.[19] When f is an order of presentation and the alternatives within each block of the partition are rearranged in or against this order – because the agent chooses the first or the last among seemingly equivalent alternatives – then the process is a successive choice from list model (see Rubinstein and Salant, 2006, for a definition). Special cases include rational choice for the vector $(k_1, k_2, \ldots, k_s) = (1, 1, \ldots, 1)$, perfect-recall satisficing for $(k, 1, \ldots, 1)$, no-recall satisficing for $(k, m_X - k)$, and situations where the welfare preference has no impact on behavior for $k_1 = m_X$.

The following result shows that there are never any dominated frames for models with the frame-cancellation property.

Proposition 2.3 *If d has the frame-cancellation property, then $M(\Lambda) = F$ for all Λ.*

If the frame-cancellation property holds, then irrespective of how many data points we have collected, we will never know enough to improve upon any given frame. According to our earlier analogy, all frames are always Pareto efficient. If we want to select between them, we need to resort to approaches that can be used to compare Pareto efficient allocations, involving stronger assumptions such as probabilistic beliefs (see Section 2.6.2).

behavior can be modelled as the maximization of some non-strict preference that is turned strict by ordering its indifference sets in or against the list order (Proposition 2, p. 8).

[19] For any $\succeq \in P$, since $\succeq \in P_{\succeq}$ holds, we have $d(\succeq, f_1) \in P_{\succeq}$ for any $f_1 \in F$. Then we also obtain $d(d(\succeq, f_1), f_2) = d(\succeq, f_2)$ for any $f_2 \in F$, which is the frame-cancellation property. We note that there are models with the frame-cancellation property that do not belong to the class of limited sensitivity models. Any model with frame-cancellation allows us to partition P into maximal indistinguishable sets of preferences, very similar to the sets $P_{\succeq}$ in the limited sensitivity model, but these sets will not in general be generated by some vector $(k_1, k_2, \ldots, k_s)$ as required by the limited sensitivity model.

2.5 Identifiable Preferences

We now turn to models with identifiable welfare preferences, which guarantee knowledge of an optimal nudge once a maximal data set has been observed. Collecting a maximal data set requires observing the agent under all m_F frames, however, which might be beyond our means. We are thus interested in optimal data gathering procedures and the required quantity of information. The idea is that a regulator, who ultimately seeks to impose the optimal nudge, is also able to impose a specific sequence of frames on the agent, with the goal of eliciting the welfare preference.

For each $s \in \{0, 1, \ldots, m_F\}$, let

$$L_s = \{\Lambda | P(\Lambda) \neq \varnothing \text{ and } |\Lambda| = s\}$$

be the collection of data sets that do not falsify the behavioral model and contain exactly s observations, i.e., observations for s different frames. In particular, $L_0 = \{\varnothing\}$, and L_{m_F} consists of all maximal data sets. Then $L = L_0 \cup L_1 \cup \ldots \cup L_{m_F-1}$ is the collection of all possible data sets except the maximal ones. An elicitation procedure dictates for each of these data sets a yet unobserved frame, under which the agent is to be observed next.

Definition 2.4 *An elicitation procedure is a mapping $e : L \to F$ with the property that, for each $\Lambda \in L$, there does not exist $(\succeq, f) \in \Lambda$ such that $e(\Lambda) = f$.*

A procedure e starts with the frame $e(\varnothing)$ and, if the welfare preference is $\succeq$, generates the first data set $\Lambda_1(e, \succeq) = \{(d(\succeq, e(\varnothing)), e(\varnothing))\}$. It then dictates the different frame $e(\Lambda_1(e, \succeq))$ and generates a larger data set $\Lambda_2(e, \succeq)$ by adding the resulting observation. This yields a sequence of expanding data sets described recursively by $\Lambda_0(e, \succeq) = \varnothing$ and

$$\Lambda_{s+1}(e, \succeq) = \Lambda_s(e, \succeq) \cup \{(d(\succeq, e(\Lambda_s(e, \succeq))), e(\Lambda_s(e, \succeq)))\},$$

until the maximal data set $\Lambda_{m_F}(e, \succeq) = \bar{\Lambda}(\succeq)$ is reached. Hence all elicitation procedures deliver the same outcome after m_F steps, but typically differ at earlier stages. A procedure does not use any exogenous information about the welfare preference, but the frame to be dictated next can depend on the information generated endogenously by the growing data set.[20]

We now define the complexity n of the nudging problem as the number of steps that the quickest elicitation procedure requires until it identifies an optimal nudge for sure. Formally, let

$$n(e, \succeq) = \min\{s \mid G(\Lambda_s(e, \succeq)) \neq \varnothing\}$$

denote the first step at which e identifies an optimal nudge if the welfare preference is $\succeq$. Since this preference is unknown, e guarantees a result only after $\max_{\succeq \in P} n(e, \succeq)$ steps. With E denoting the set of all elicitation procedures, we have to be prepared to gather

$$n = \min_{e \in E} \max_{\succeq \in P} n(e, \succeq)$$

[20] Notice that an elicitation procedure dictates frames also for pre-collected data sets that itself never generates. We tolerate this redundancy because otherwise definitions and proofs would become substantially more complicated, at no gain.

data points before we can nudge successfully.

To illustrate the concepts, we first consider the limited search model (assuming $m_X \geq 3$ to make all preferences identifiable). The following result shows that learning and nudging are relatively simple in this model.

Proposition 2.4 *For any $m_X \geq 3$, the limited search model satisfies*

$$n = \begin{cases} 3 & \text{if } k = m_X/2 \text{ and } k \text{ is odd,} \\ 2 & \text{otherwise.} \end{cases}$$

To understand our construction of an optimal elicitation procedure for the limited search model, consider again Example 2. The procedure starts with an arbitrary frame, $f = \{a, b\}$, and generates the behavioral preference $a \succ_3 b \succ_3 c \succ_3 d$. We now know that the welfare preference satisfies $a \succ b$ and $c \succ d$. The second frame is constructed by taking the top element from f and the bottom element from $X \backslash f$, which yields $f' = \{a, d\}$. From the induced behavioral preference $a \succ_7 d \succ_7 c \succ_7 b$ we learn that $a \succ d$ and $c \succ b$. This information is enough to deduce that a and c are the two welfare-optimal alternatives, because both b and d are worse than each of them. If instead at the second step we had learned that $a \succ d$ and $b \succ c$, we could have concluded that a and b are optimal. If we had learned that $d \succ a$, we could have concluded that c and d are optimal. This argument can be generalized. If $k = m_X/2$ and k is even, for instance, the second frame is constructed to contain the $k/2$ best alternatives from the previous first result page and the $k/2$ worst alternatives from the previous second result page. It can be shown that the k welfare-best alternatives can always be deduced from the resulting data set.

The nudging complexity is surprisingly small for the limited search model. This begs the question to what extent it is representative for more general models. It obviously always holds that $n \leq m_F$ if all welfare preferences are identifiable, but the number of frames m_F can be extremely large (see footnote 15). We therefore derive a tighter bound on n next. The result rests on the insight that there is always an elicitation procedure that guarantees a reduction of the set of possible welfare preferences at each step. Since there are $m_P(m_X) = m_X!$ different welfare preferences that the agent might have ex ante, an elicitation procedure that reduces the set of possible preferences at each step guarantees identification of the preference and the optimal nudge after at most $m_X! - 1$ steps.

Proposition 2.5 *Any behavioral model with identifiable preferences satisfies $n \leq m_X! - 1$.*

It turns out that the bound presented in Proposition 2.5 is tight, because there are models for which it is reached. We illustrate this with the following model, which describes an, admittedly, strong effect of framing.

Model 4 (Strong Priming). The framing of the decision problem suggests that there is a unique proper way of deciding (e.g. priming, persuasion, demand effects). Formally, a frame $f \in F = P$ is identified with the preference that it conveys as being the proper behavior. The effect of the frame is strong, in the sense that the agent can be manipulated to behave in the suggested way whenever there is at least some agreement between the suggestion and the welfare preference. Manipulation fails only when the agent's welfare preference is exactly opposite of

the suggestion. In this case the agent behaves in an arbitrarily distorted way that uniquely identifies him. For any $\succeq \in P$, let $o(\succeq)$ denote the opposite order of $\succeq$. Assume $m_X \geq 3$ and let $b : P \to P$ be a bijective mapping such that $b(\succeq) \notin \{\succeq, o(\succeq)\}$, for all $\succeq \in P$. Then

$$d(\succeq, f) = \begin{cases} f & \text{if } f \neq o(\succeq), \\ b(\succeq) & \text{if } f = o(\succeq). \end{cases}$$

Proposition 2.6 *The strong priming model satisfies* $n = m_X! - 1$.

In the strong priming model, identification of the optimal nudge actually requires identification of the welfare preference, because each frame is optimal only for exactly one welfare preference. This takes all $m_X! - 1$ steps, because observation of behavior under a frame either reveals a specific welfare preference to be the true one, or it excludes it from the set of possible welfare preferences. No matter in which order frames are dictated by the elicitation procedure, it is always possible that the agent's welfare preference is the one not revealed until the end. Hence, learning is particularly slow in this model.

Taken together, Propositions 2.5 and 2.6 are bad news for nudging. The tight bound on n grows more than exponentially in the number of alternatives. Thus, nudging may quickly become infeasible despite the general identifiability of preferences.

2.6 Nudging with Additional Information

2.6.1 Restricted Domains

Throughout the previous analysis we have maintained the assumption that the regulator considers all preferences over the set of alternatives feasible. In some situations, however, the regulator may be able to rule out certain preferences beforehand using non-choice information. For instance, criteria such as non-satiation or an agreement with some objective dimension of the alternatives may sometimes be uncontroversial and reduce the set of plausible preferences. We can model situations in which some welfare preferences are excluded from the outset by restricting the domain of preferences to some non-empty $\tilde{P} \subseteq P$. We then replace the set $P(\Lambda)$ of welfare preferences that are consistent with data set Λ by $\tilde{P}(\Lambda) = P(\Lambda) \cap \tilde{P}$. Based on this modified definition, all further concepts remain unchanged. We will explore two different implications of such domain restrictions. First, models with non-identifiable preferences may become identifiable. Second, the complexity of the elicitation procedure can be reduced.

We extend Definition 2.2 by saying that a preference $\succeq \in \tilde{P}$ is identifiable on $\tilde{P}$ if for each $\succeq' \in \tilde{P}$ with $\succeq' \neq \succeq$, there exists $f \in F$ such that $d(\succeq, f) \neq d(\succeq', f)$. It then follows exactly as for Proposition 2.1 that $G(\bar{\Lambda}(\succeq))$ is non-empty if and only if $\succeq$ is identifiable on $\tilde{P}$. Hence we will call $\tilde{P}$ a *nudging domain* if all its elements are identifiable on $\tilde{P}$. The universal domain P is a nudging domain if and only if all welfare preferences are identifiable as defined previously. To characterize nudging domains more generally, let

$$P_{\succeq} = \{\succeq' \in P \,|\, d(\succeq', f) = d(\succeq, f), \forall f \in F\}$$

be the equivalence class of welfare preferences that are indistinguishable from $\succeq$, and denote by $\bar{P} = \{P_{\succeq} \mid \succeq \in P\}$ the set of all these equivalence classes, which form a partition of P. Then it follows that $\tilde{P}$ is a nudging domain if and only if $|\tilde{P} \cap P_{\succeq}| \leq 1$ for all $\succeq \in P$, i.e., the domain $\tilde{P}$ can contain at most one element from each of the behaviorally equivalent classes of preferences.

Unfortunately, this may not be a particularly plausible or easily justifiable requirement. Consider the perfect-recall satisficing model. The set $P_{\succeq}$ contains all preferences which agree with $\succeq$ with respect to the bottom $m_X - k$ alternatives and their ranking. Hence, the restriction necessary to obtain identifiable preferences is that for each selection and ordering of the bottom $m_X - k$ alternatives, there exists at most one preference in $\tilde{P}$. Put differently, the preference over the bottom alternatives must fully determine the preference over all alternatives. This is very different from often studied domain restrictions such as single-peaked preferences (which do not constitute a nudging domain for the satisficing model or any of the other models with non-identifiable preferences studied in this paper).

The extent to which the domain needs to be restricted can still be interpreted as a measure of the degree of non-identifiability of a model. Different models may be unambiguously comparable by their demand for exogenous information. We show this for the satisficing models with perfect recall and no recall.

Proposition 2.7 *Any nudging domain for no-recall satisficing is also a nudging domain for perfect-recall satisficing, while the converse is not true whenever* $k < m_X - 1$.

A satisficer with no recall is harder to nudge than a satisficer with perfect recall (whenever the two are behaviorally different), because more knowledge about the welfare preference is necessary and less can be learned from behavior. As a general rule, a model comparison as in Proposition 2.7 is possible whenever the partition $\bar{P}$ is finer for one model than for another.

Let us now consider the effect of domain restrictions on the complexity of nudging. Whenever $\tilde{P}$ is a nudging domain for model d, we can adapt our previous definition of complexity to

$$\tilde{n} = \min_{e \in E} \max_{\succeq \in \tilde{P}} n(e, \succeq),$$

no matter whether or not d has identifiable preferences on the universal domain P. We obtain the following generalization of Propositions 2.5 and 2.6, which shows that the modified complexity bound naturally mirrors the amount of non-choice information we put in.

Proposition 2.8 *Any behavioral model on a nudging domain* $\tilde{P}$ *satisfies* $\tilde{n} \leq |\tilde{P}| - 1$. *The strong priming model satisfies* $\tilde{n} = |\tilde{P}| - 1$.

2.6.2 Probabilistic Beliefs

We now explore the possibility to introduce non-choice-based prior information in the form of probabilistic beliefs. We assume that the regulator has a prior belief p over the set of welfare preferences P. We number the preferences in the order of their prior probabilities, so that $P = \{\succeq_1, \succeq_2, \ldots, \succeq_{m_X!}\}$ with $p_1 \geq p_2 \geq \ldots \geq p_{m_X!} > 0$.[21] Beliefs can be utilized in different

[21] In contrast to the previous subsection, here we make the full support assumption $p_{m_X!} > 0$. This is for simplicity and allows us to circumvent technical issues with Bayesian updating which would otherwise require a redefinition of elicitation procedures.

ways. A first possibility is to replace our previous notion of complexity n by the expected complexity

$$\bar{n} = \min_{e \in E} \sum_{i=1}^{m_X!} p_i n(e, \succeq_i).$$

While n was the minimal number of observations necessary to guarantee identification of an optimal nudge, $\bar{n}$ can be thought of as the average running time of the quickest elicitation procedure. Different procedures may be required to achieve n or $\bar{n}$, respectively. The following result provides the expected complexity for the strong priming model, which we used before to illustrate the potentially large complexity of the elicitation problem.

Proposition 2.9 *The strong priming model satisfies*

$$\bar{n} = \sum_{i=1}^{m_X!-1} p_i i + p_{m_X!}(m_X! - 1).$$

At any given step, the elicitation procedure that has not yet identified the optimal nudge should always try to verify or exclude the remaining welfare preference with the *highest* belief probability, by prescribing the frame that corresponds to the opposite of this preference. The elicitation process then concludes with highest possible probability at every step, which gives rise to the formula in the proposition.

The complexity $\bar{n}$ and its behavior for large m_X depend on the shape of prior beliefs. As an example of a relatively informative prior, consider a (truncated) geometric distribution where the prior probabilities are given by

$$p_i = \rho^{i-1} \left(\frac{1-\rho}{1-\rho^{m_X!}} \right)$$

for some parameter $\rho \in (0,1)$. In Appendix B.2 we show that $\lim_{m_X \to \infty} \bar{n} = 1/(1-\rho)$ holds for this distribution. The expected complexity thus remains bounded as the number of alternatives grows, and it may be small if ρ is small. On the other hand, for a uniform prior, where $p_i = 1/m_X!$, we show that $\bar{n}$ is still of the same order of magnitude as the previous $n = m_X! - 1$, and grows more than exponentially in the number of alternatives.

Let us therefore consider a second way in which belief-dependent complexity could be defined. In particular, we introduce a probabilistic notion of optimality of a nudge. Let $\pi_\Lambda(\succeq)$ denote the updated belief probability that the regulator attaches to welfare preference $\succeq$ when the data set Λ has been collected. We thus have $\pi_\varnothing(\succeq_i) = p_i$ and can apply Bayesian updating to obtain

$$\pi_\Lambda(\succeq) = \begin{cases} \pi_\varnothing(\succeq) / \left(\sum_{\succeq' \in P(\Lambda)} \pi_\varnothing(\succeq') \right) & \text{if } \succeq \in P(\Lambda), \\ 0 & \text{otherwise,} \end{cases}$$

for all data sets Λ with $P(\Lambda) \neq \varnothing$. In addition to narrowing down the set of possible welfare preferences, collecting data magnifies differences in prior beliefs on $P(\Lambda)$. Now let $\varphi_\Lambda(f)$ denote the probability that frame f is an optimal nudge. With the definition of $P(\Lambda, f) = \{\succeq \in P(\Lambda) \mid$

$d(\succeq, f) = \succeq\}$ as the set of remaining preferences for which f is non-distorting, we can calculate

$$\varphi_\Lambda(f) = \sum_{\succeq \in P(\Lambda, f)} \pi_\Lambda(\succeq).$$

We will denote by $\bar{\varphi}_\Lambda = \max_{f \in F} \varphi_\Lambda(f)$ the confidence that an optimally chosen frame induces non-distorted behavior.

From our previous arguments we obtain that $\varphi_\Lambda(f) = 1$ if and only if $f \in G(\Lambda)$. Hence the complexity n was based on the requirement that we want to ensure complete confidence, $\bar{\varphi}_\Lambda = 1$. We may now content ourselves with identifying a frame that is optimal with a sufficiently large probability $q \in (0, 1]$. The optimal elicitation procedure then is the one that guarantees a level $\bar{\varphi}_\Lambda \geq q$ as quickly as possible. This is captured by the generalized definition $n(q, e, \succeq) = \min\{s \mid \bar{\varphi}_{\Lambda_s(e,\succeq)} \geq q\}$ and

$$n(q) = \min_{e \in E} \max_{\succeq \in P} n(q, e, \succeq).$$

The following result provides the generalized complexity for the strong priming model.

Proposition 2.10 *The strong priming model satisfies that $n(q)$ is the smallest integer $k \geq 0$ for which*

$$\sum_{j=1+k}^{m_X!-1} p_{j+1} \leq p_1 \left(\frac{1-q}{q} \right).$$

At any given step, a generalized optimal procedure that has not yet identified the optimal nudge should always try to verify or exclude the remaining welfare preference with the *second-highest* belief probability, by prescribing the frame that corresponds to the opposite of this preference. It can always occur that the procedure still does not identify the welfare preference, but in that case it guarantees maximal posterior beliefs.

The result implies our previous result for n when we consider the limit as $q \to 1$, i.e., for large enough q we always obtain $n(q) = m_X! - 1$. With a uniform prior, for instance, we can rearrange the condition in Proposition 2.10 to obtain the ceiling

$$n(q) = \max \left\{ \left\lceil (m_X! - 1) - \left(\frac{1-q}{q} \right) \right\rceil, 0 \right\},$$

as shown in Appendix B.2. This implies $n(q) = m_X! - 1$ whenever $q > 1/2$. The uniform prior can be interpreted as the criterion of counting the welfare preferences for which a given frame is optimal. The result thus shows that the previous complexity bound remains the same as long as we require optimality for a strict majority of welfare preferences. On the other hand, the combination of an informative prior belief and a low degree of required confidence can reduce the complexity of the nudging problem substantially. An extreme case is $p_1 \geq q$, so that the prior belief already provides sufficient confidence and $n(q) = 0$ follows. For the geometric distribution, we show that the generalized complexity remains bounded as the number of alternatives grows whenever $q < 1$.

The complexity $n(q)$ is also interesting for models with non-identifiable preferences. Our previous results imply that we can never achieve $\bar{\varphi}_\Lambda = 1$ for such models, but we may be able to achieve $\bar{\varphi}_\Lambda \geq q$ when q is sufficiently small. We can easily extend our definition of $n(q)$ to the

case of non-identifiable preferences, by defining $n(q, e, \succeq) = \infty$ whenever elicitation procedure e never achieves a confidence of q or larger when $\succeq$ is the true welfare preference, i.e., when

$$q > \overline{q}(e, \succeq) = \max_{s \in \{0,\ldots,m_F\}} \bar{\varphi}_{\Lambda_s(e,\succeq)}.$$

Let $\underline{q} = \bar{\varphi}_{\varnothing}$ denote the prior confidence and $\overline{q} = \max_{e \in E} \min_{\succeq \in P} \overline{q}(e, \succeq)$ the maximal confidence that an optimal procedure can guarantee. We thus have $0 < \underline{q} \leq \overline{q} < 1$ and

$$n(q) \in \begin{cases} \{0\} & \text{if } q \leq \underline{q}, \\ \{1, \ldots, m_X! - 1\} & \text{if } \underline{q} < q \leq \overline{q}, \\ \{\infty\} & \text{if } \overline{q} < q. \end{cases}$$

The fact that $n(q) \leq m_X! - 1$ when $\underline{q} < q \leq \overline{q}$ follows as for Proposition 2.5.[22] For models with the frame-cancellation property, we can make the following more precise statement.

Proposition 2.11 *If d has the frame-cancellation property, then*

$$n(q) = \begin{cases} 0 & \textit{if } q \leq \underline{q}, \\ 1 & \textit{if } \underline{q} < q \leq \overline{q}, \\ \infty & \textit{if } \overline{q} < q. \end{cases}$$

As shown previously, models with the frame-cancellation property stand out because the scope of learning about optimal nudges is particularly limited. Proposition 2.11 shows that, at the same time, the speed of learning is particularly fast for these models. All that can be learned under the frame-cancellation property is learned already after a single observation. Hence if we are willing to reduce our confidence aspiration, models with the frame-cancellation property stand out because of the simplicity of learning.

We conclude by pointing out two potential pitfalls of this result. First, significant learning will only be possible if there is significant information already in the prior beliefs. Second, achieving a satisfactory level of confidence is not tantamount to having a good guidance in the choice between frames. We illustrate this by studying the case of a uniform prior distribution. Let $\bar{s}$ denote the average size of the elements of partition $\bar{P}$, i.e., the average number of preferences in an indistinguishable equivalence class.

Proposition 2.12 *If d has the frame-cancellation property and prior beliefs are uniform, then $\underline{q} = \overline{q} = 1/\bar{s}$. Furthermore, $\varphi_\Lambda(f) = \varphi_\Lambda(f')$ for all $f, f' \in F$ and all Λ.*

With a uniform prior, no procedure guarantees to increase the prior confidence. In addition, all frames are always equally likely to be optimal. A regulator can thus never do better than by choosing a frame at random.

[22] There is always an elicitation procedure that guarantees a reduction of the set of possible welfare preferences at each of the initial steps, until a further reduction is no longer possible at all. Thus the entire range of achievable confidence levels can be achieved during the first $m_X! - 1$ steps.

2.7 Nudging with Limited Information

2.7.1 Model Uncertainty

We have so far assumed that there is a unique conjecture about the behavioral model, while it may be more appropriate to assume that the regulator considers a number of different models possible. We can replace the assumption of a unique behavioral model by the assumption that the regulator considers any distortion function $d \in D$ possible, where D is a given set of conjectures. For instance, there could be uncertainty about the aspiration level of a satisficer, or one of the models in D could be the rational agent.[23] As a consequence, we no longer have to learn about the welfare preference only, but about the pair $(d, \succeq) \in D \times P$ of the distortion function and the welfare preference.[24]

Let $\bar{\Lambda}(d, \succeq) = \{(d(\succeq, f), f) \mid f \in F\}$ denote the maximal data set generated by the pair $(d, \succeq)$. Then the set of pairs $(d, \succeq)$ that are consistent with data set Λ is $DP(\Lambda) = \{(d, \succeq) \mid \Lambda \subseteq \bar{\Lambda}(d, \succeq)\}$. We again assume that $DP(\Lambda)$ is non-empty, i.e., at least one conjecture is not falsified by the data. Once we have narrowed down the set of model-preference pairs to $DP(\Lambda)$, we obtain the equivalence class of frame f by $[f]_\Lambda = \{f' \mid d(\succeq, f) = d(\succeq, f'),\ \forall (d, \succeq) \in DP(\Lambda)\}$. We can then modify our definition of the binary nudging relation in a natural way, taking into account that both model and welfare preference are unknown. In particular, we define $[f]_\Lambda\ N(\Lambda)\ [f']_\Lambda$ if for each $(d, \succeq) \in DP(\Lambda)$ it holds that $c(d(\succeq, f), S) \succeq c(d(\succeq, f'), S)$ for all non-empty $S \subseteq X$, so that for each remaining behavioral model the agent's choice under frame f is at least as good as under f', no matter which of the welfare preferences that are consistent with the behavioral model and the data set is the true one.

We are again interested in the existence of an optimal nudge. By the same reasoning as in Section 2.3, we consider maximal data sets only. An immediate extension of Definition 2.2 could require identifiability of $\succeq$ in d, for a given pair $(d, \succeq)$. This property is in fact necessary but no longer sufficient for the existence of an optimal nudge. It rules out that the maximal data set $\bar{\Lambda}(d, \succeq)$ could have been generated by a different welfare preference $\succeq'$ and the same model d, but it does not rule out that it could have been generated by a different welfare preference $\succeq'$ and a different model d'. Since two behaviorally equivalent model-preference pairs $(d, \succeq)$ and $(d', \succeq')$ can have different normative implications (see e.g. Bernheim, 2009; Kőszegi and Rabin, 2008b; Masatlioglu et al., 2012), identifiability in the extended setting must aim at all aspects of the pair $(d, \succeq)$ that are normatively relevant.

Definition 2.5 *Pair $(d, \succeq)$ is virtually identifiable if for each $(d', \succeq') \in D \times P$ with $\succeq' \neq \succeq$, there exists $f \in F$ such that $d(\succeq, f) \neq d'(\succeq', f)$.*

Virtual identifiability implies that the welfare preference $\succeq$ is known for sure once the maximal data set has been collected. It still allows for some uncertainty about the behavioral model, but only to the extent that we might not be able to predict the behavior of an agent with a different welfare preference $\succeq' \neq \succeq$.

[23] It is central to the idea of asymmetric paternalism (Camerer et al., 2003) that there are different types of agents, some of which are rational and should not be restricted by regulation.

[24] We continue to assume that there is a non-distorting frame for each pair $(d, \succeq)$, which will typically depend both on the model and on the welfare preference.

Proposition 2.13 *With model uncertainty, $G(\bar{\Lambda}(d, \succeq))$ is non-empty if and only if $(d, \succeq)$ is virtually identifiable.*

We can have multiple models with identifiable preferences each, that, if considered jointly, do not have virtually identifiable model-preference pairs. Model uncertainty of this type poses a fundamental problem to nudging. On the other hand, adding a rational agent to any given behavioral model with identifiable preferences preserves the property of virtually identifiable model-preference pairs. Thus the possibility of agents being rational has no substantial impact on our previous results.

The analysis in Sections 2.4 and 2.5 could also be adapted to the case of model uncertainty. For instance, if each distortion function $d \in D$ satisfies the frame-cancellation property, then it follows immediately that no data set allows us to exclude any dominated frame. Applications include the uncertainty about a satisficer's aspiration level. With virtually identifiable model-preference pairs, on the other hand, elicitation procedures now generate sequences of expanding data sets with the goal of learning about both preferences and models.

We could go one step further and dispense with any model conjecture. Instead of following our model-based approach to behavioral welfare economics, we could work with the purely choice-based approach by Bernheim and Rangel (2009).[25] In fact, we can easily adapt our definition of the binary nudging relation and evaluate the frame-induced choices based on the weak unambiguous choice relation R' (Bernheim and Rangel, 2009, p. 60), rather than on a set of welfare preferences. Formally, a generalized choice situation (GCS) consists of a set of alternatives $S \subseteq X$ and a frame $f \in F$, and a choice correspondence describes the chosen alternatives for each GCS that we have observed. Let us assume that the observed choice has always been a unique alternative $C(S, f) \in S$. To eliminate all traces of non-choice-based theories about mistakes, let us also assume that all the observed GCSs are welfare-relevant. Now consider two frames f and f' of which we know that they have a differential impact on behavior, i.e., we have observed two GCSs $(\bar{S}, f)$ and $(\bar{S}, f')$ with $C(\bar{S}, f) = x \neq y = C(\bar{S}, f')$. In line with our previous analysis, we could say that f is a weak unambiguous nudge over f' if $C(S, f)\, R'\, C(S, f')$ holds for all matching pairs (S, f) and (S, f') that we have observed. It follows immediately from the definition of R' that such a ranking is impossible. The mere fact that $C(\bar{S}, f) = x \neq y = C(\bar{S}, f')$ implies that neither $xR'y$ nor $yR'x$ holds, and hence neither of the two frames can be a weak unambiguous nudge over the other. Nudging is impossible without assumptions about decision mistakes (as already pointed out by Bernheim and Rangel, 2009, p. 62).

2.7.2 Imperfectly Observable Frames

So far we have assumed that frames are perfectly observable and controllable by the regulator. Since a frame can be very complex, this assumption deserves to be relaxed. The generalization also allows us to model fluctuating internal states of the agent that affect her choices. For

[25] Another interesting choice-based approach is due to Apesteguia and Ballester (2015), who propose using as a welfare benchmark the preference that is closest to a given behavior, measured by their "swaps" criterion. Their framework does not allow for frames, but it would be interesting to develop the respective generalization and derive the implications for nudging.

instance, consider a modified satisficing model in which the aspiration level k fluctuates in a non-systematic and unobservable way, as in the original RS model. We can capture this by including the aspiration level into the description of the frame (k affects choice but not welfare), but the extended frame cannot be fully observable and controllable for an outsider.

Imperfect observability can be modelled as a structure $\Phi \subseteq 2^F$ with the property that for each $f \in F$ there exists $\phi \in \Phi$ with $f \in \phi$. The interpretation is that the regulator observes only sets of frames $\phi \in \Phi$ and does not know under which of the frames $f \in \phi$ the agent was acting. The example with a fluctuating aspiration level can be modelled as $F = P \times \{2, \dots, m_X\}$ and $\Phi = \{\phi_p \mid p \in P\}$ for $\phi_p = \{(p,k) \mid k \in \{2, \dots, m_X\}\}$. A behavioral data set is a subset $\Lambda \subseteq P \times \Phi$, where $(\succeq', \phi') \in \Lambda$ means that the agent has been observed behaving according to $\succeq'$ when the frame must have been one of the elements of ϕ'. Thus a welfare preference $\succeq$ is consistent with Λ if for each $(\succeq', \phi') \in \Lambda$ we have $\succeq' = d(\succeq, f')$ for some $f' \in \phi'$, so that $\succeq$ might have generated the data set from the regulator's perspective. The set of welfare preferences that are consistent with Λ is $P(\Lambda) = \{\succeq \mid \Lambda \subseteq \bar{\Lambda}(\succeq)\}$, where $\bar{\Lambda}(\succeq) = \{(d(\succeq, f), \phi) \mid f \in \phi \in \Phi\}$ is again the maximal data set for $\succeq$. Note that a non-singleton set of frames ϕ can appear more than once in a maximal data set, combined with different behavioral preferences. This also implies that the cardinality of $\bar{\Lambda}(\succeq)$ is no longer the same for all $\succeq \in P$, because two different frames $f, f' \in \phi$ might generate two different observations for some preference but only one observation for another preference.

In many applications, such as a satisficing model with fluctuating aspiration level, it is reasonable to assume that the same Φ applies to observing and nudging, i.e., the frame dimensions that the regulator can observe are identical to those that he can control. We allow for the more general case where a set of frames can be chosen as a nudge from a potentially different structure Φ_N.[26] When comparing two elements $\phi, \phi' \in \Phi_N$, we will not necessarily want to compare the agents' choices under each $f \in \phi$ with her choices under each $f' \in \phi'$. For instance, we want to compare orders of presentation for each aspiration level separately, not across aspiration levels. To this end, we introduce a set H of selection functions, which are functions $h : \Phi_N \to F$ with the property that $h(\phi) \in \phi$. The elements of H capture the comparisons that we need to make: when comparing ϕ with ϕ' we compare only the choices under the frames $h(\phi)$ and $h(\phi')$, for each $h \in H$. In the satisficing model we would have one $h_k \in H$ for each aspiration level $k \in \{2, \dots, m_X\}$, defined by $h_k(\phi_p) = (p,k)$. The only assumption that we impose on H is that for each $f \in \phi \in \Phi_N$ there exist $h \in H$ such that $h(\phi) = f$. We can then define the equivalence class $[\phi]_\Lambda = \{\phi' \mid d(\succeq, h(\phi')) = d(\succeq, h(\phi)), \forall (h, \succeq) \in H \times P(\Lambda)\}$ for any Λ and ϕ. As before, let $[\phi]_\Lambda N(\Lambda) [\phi']_\Lambda$ if for each $(h, \succeq) \in H \times P(\Lambda)$ it holds that $c(d(\succeq, h(\phi)), S) \succeq c(d(\succeq, h(\phi')), S)$, for all non-empty $S \subseteq X$.

Let $G(\Lambda) = \{\phi \mid [\phi]_\Lambda N(\Lambda) [\phi']_\Lambda, \forall \phi' \in \Phi_N\}$ be the set of optimal nudges. We again consider maximal data sets. An immediate extension of identifiability of $\succeq$ (Definition 2.2) could require that for each $\succeq' \neq \succeq$ there exists $f \in \phi \in \Phi$ such that $d(\succeq, f) \neq d(\succeq', f)$. This property turns

[26] In continuation of our previous approach, we assume that for each $\succeq \in P$ there exists $\phi \in \Phi_N$ such that $d(\succeq, f) = \succeq$ for all $f \in \phi$. This implies that nudging is not per se impeded by the lack of control over frames. The assumption is clearly much stronger here than before. For instance, it holds in the described satisficing application when there is perfect recall (because the order of presentation that coincides with the welfare preference is non-distorting for all possible aspiration levels) but would not hold with no recall (because the non-distorting order of presentation then depends on the aspiration level).

out to be necessary but not sufficient for $G(\bar{\Lambda}(\succeq))$ to be non-empty. It implies that the maximal data set for $\succeq$ is different from the maximal data set for every other preference, so that $\succeq$ is identified once $\bar{\Lambda}(\succeq)$ has been collected and once it is known that this set is indeed maximal. Unfortunately, the cardinality of $\bar{\Lambda}(\succeq)$ no longer carries that kind of information, as we could have $\bar{\Lambda}(\succeq) \subset \bar{\Lambda}(\succeq')$ for some $\succeq' \neq \succeq$. Upon observing $\bar{\Lambda}(\succeq)$ we then never know if we have already arrived at the maximal data set for $\succeq$, or if there is an additional observation yet to be made. Our notion of identifiability in the setting with imperfectly observable frames must therefore ensure that the maximal data set reveals itself as maximal.

Definition 2.6 *Preference $\succeq$ is potentially identifiable if for each $\succeq' \in P$ with $\succeq' \neq \succeq$, there exist $f \in \phi \in \Phi$ such that $d(\succeq, f) \neq d(\succeq', f')$ for all $f' \in \phi$.*

When frames are not directly observed, identifiability requires more than the existence of a frame $f \in \phi \in \Phi$ that distinguishes between $\succeq$ and $\succeq'$. We can exclude welfare preference $\succeq'$ as a candidate only if the observed distorted preference $d(\succeq, f)$ could not as well have been generated by $\succeq'$ for any other $f' \in \phi$. For instance, no preference is potentially identifiable in the perfect-recall satisficing model with fluctuating aspiration level.[27]

Proposition 2.14 *With imperfectly observable frames, $G(\bar{\Lambda}(\succeq))$ is non-empty if and only if $\succeq$ is potentially identifiable.*

We use the term potential identifiability because there is no guarantee that we will ever arrive at $\bar{\Lambda}(\succeq)$. An appropriately redefined elicitation procedure might impose a set of frames ϕ multiple times on the agent, but a specific element $f \in \phi$ still does not materialize. This is in contrast to the case of observable frames, where a maximal data set can always be collected in exactly m_F steps.

2.8 Savings Application

We now study an extended application of our approach to a savings problem. The question how to encourage savings has received much attention in the nudging literature from the beginning.[28] The application also extends our previous setting in various directions, so it illustrates the flexibility and portability of our approach.

We consider a two-period environment with alternatives $x = (x_1, x_2) \in X = \mathbb{R}^2_+$ that specify a present payment of x_1 and a future payment of x_2. In line with much of the literature that estimates discount rates from behavioral data (see e.g. Cohen, Ericson, Laibson, and White, 2016), we assume that the agent is risk-neutral. This is an acceptable approximation when the agent's background consumption is large relative to the payoff consequences of the considered choices. We thus focus on the restricted domain $\tilde{P}$ of welfare preferences that can be represented by a utility function of the form

$$u(x_1, x_2) = x_1 + \delta x_2,$$

[27] To see why, note that two preferences which coincide except for the ranking of the two top alternatives are behaviorally equivalent for every order of presentation and every aspiration level $k \geq 2$. This is different if we allow the agent to be sometimes rational ($k = 1$) as in the original RS model, in which case all preferences are potentially identifiable.

[28] For a recent contribution see Bernheim, Popov, and Fradkin (2015), who derive weak generalized Pareto optimal 401(k) defaults in the sense of Bernheim and Rangel (2009), with and without pruning.

for some unknown discount factor $0 < \delta \leq 1$. This environment extends our previous setting by allowing an infinite set of alternatives and non-strict preferences.

We aim at modelling that the framing of the decision problem necessarily prompts the agent to take a more *short-run* or a more *long-run* perspective. Examples of frames that induced more or less patient choices include the timing of the choice, the default allocation, the status quo, the phrasing of the question, or whether the agent is hungry or sated when making the choice (see e.g. Loewenstein and Prelec, 1992). The literature has described various channels through which these frames operate; they may focus the agent's attention on a particular time period, activate hot or cold states, moderate visceral influences, rouse different behavioral selves, or set reference points. We capture these effects by defining two abstract frames, a short-run frame f_S and a long-run frame f_L. We do not advocate the simple view that the choices under one of the two frames are always welfare-maximizing. Frame f_S may induce "lapses of self-control" but frame f_L may cause future benefits to be "excessively intellectualized at arm's length" (Bernheim and Rangel, 2009, p. 58). Hence we believe it is most reasonable to assume that, if anything, the choices under f_S are present-biased while the choices under f_L are future-biased. We model this by assuming that an agent with true discount factor δ acts as if maximizing the preference represented by the utility function

$$u_S(x_1, x_2) = \gamma x_1 + \delta x_2$$

under frame f_S, and the preference represented by the utility function

$$u_L(x_1, x_2) = x_1 + \gamma\delta x_2$$

under frame f_L. The parameter $\gamma \geq 1$ captures each frame's distortion towards one of the two time periods, and its magnitude measures the agent's susceptibility to framing. We assume that γ is an unknown parameter of the behavioral model and treat (γ, δ) as the model-preference pair that has to be elicited from the behavioral data set. Since both frames are distorting whenever $\gamma > 1$, here we are relaxing the previous assumption that a non-distorting frame must exist for each model-preference pair.

We first turn to the problem of eliciting (γ, δ). After a normalization of u_S it follows that the agent applies the behavioral discount factor $\delta_S = \delta/\gamma$ under frame f_S. Holding the frame fixed, this discount factor can be measured by observing the agent's choices from sufficiently many different subsets $S \subseteq X$. To that end, the experimental literature that we will discuss below typically uses paradigms such as multiple price lists or matching (Cohen et al., 2016). There are still many model-preference pairs that are consistent with some measured δ_S (except if $\delta_S = 1$), but the procedure can be repeated to also obtain a measure of the behavioral discount factor $\delta_L = \delta\gamma$ under frame f_L. The maximal data set $\Lambda = \{(\delta_S, f_S), (\delta_L, f_L)\}$ then reveals that (γ, δ) is given by $\gamma = \sqrt{\delta_L/\delta_S}$ and $\delta = \sqrt{\delta_L\delta_S}$. Hence each model-preference pair is (fully) identifiable.

We now turn to the problem of nudging an agent who is characterized by (γ, δ). If we required one frame to outperform the other frame for all (compact) choice sets $S \subseteq X$, as we did in our previous analysis where a non-distorting frame was always available, we would

immediately find that none of the frames is a successful nudge over the other.[29] We therefore work with the weaker but reasonable requirement that only the choices in a prespecified range of plausible market conditions have to be improved by the optimal nudge. A market condition is described by the interest rate r and the present value y of the agent's investment opportunities, which jointly generate a budget line $X(r,y) = \{x \in X \mid x_1 + x_2/(1+r) = y\}$. Let C be a set of market conditions (r,y) for which we want to ensure optimal choices. We then say that frame f_S is a weakly successful nudge over frame f_L if each u_S-optimal element in choice set S is weakly u-better than each u_L-optimal element in S, and this holds for all compact subsets $S \subseteq X(r,y)$ for all $(r,y) \in C$. Intuitively, we ensure that the agent's choices under frame f_S are welfare-better than her choices under frame f_L for all admissible market conditions. We consider subsets $S \subseteq X(r,y)$ instead of only the entire budget lines $X(r,y)$ to reflect the possibility that there are floors or caps on investment, or that savings rates must be selected from a finite set. The definition of f_L being a nudge over f_S is analogous.

The set C may be generated by an interval of interest rates that we deem plausible, and a range of money amounts potentially available for investment to the agent (e.g. shares of current labor income). In general we only assume that C is compact and connected but does not necessarily have a product structure. Let $\underline{r}$ denote the smallest interest rate and $\overline{r}$ the largest interest rate in C. We allow interest rates to be negative but assume that $\underline{r} > -1$. Then we obtain the following result on the possibility of nudging.

Proposition 2.15 *Frame f_S is an optimal nudge if $\delta \leq 1/(1+\overline{r})$. Frame f_L is an optimal nudge if $1/(1+\underline{r}) \leq \delta$. Both frames are undominated if $1/(1+\overline{r}) < \delta < 1/(1+\underline{r})$.*

Impatient agents are nudged optimally by the short-run frame, and patient agents are nudged optimally by the long-run frame. The precise thresholds for δ depend on the range of market interest rates for which the nudge is supposed to work. The thresholds do not depend on the size of the agent's susceptibility to framing γ, and therefore the behavior of the nudgeable types is far from being unambiguous. For instance, an impatient agent with $\delta < 1/(1+\overline{r})$ may behave very patiently and choose the highest available savings rate under frame f_L. This happens for sufficiently high interest rates whenever $\gamma\delta > 1/(1+\overline{r})$, and it even happens for all interest rates when $\gamma\delta > 1/(1+\underline{r})$. Our analysis then recommends to overrule these seemingly cold and rational long-run choices, by nudging the agent in a way that induces more impatient behavior. Conversely, our analysis recommends to correct lapses of self-control of patient agents, by nudging them to take the long-run perspective.

We now use empirical estimates of behavioral discount factors to obtain quantitative predictions from our model. As discussed above, several of the behavioral anomalies in intertemporal choice could be mapped into the model, because they can be understood as a frame-driven conflict between the short-run and the long-run perspective. A particularly lucid effect is the asymmetry between *delay* and *speed-up* framing (Benzion, Rapoport, and Yagil, 1989; Loewenstein, 1988; Shelley, 1993; Weber, Johnson, Milch, Chang, Brodscholl, and Goldstein, 2007).

[29] Just consider the binary choice set $S = \{x, x'\}$ where $x = (x_1, x_2)$ and $x' = (x_1 - \epsilon_1, x_2 + \epsilon_2)$. When $\epsilon_2\delta/\gamma < \epsilon_1 < \epsilon_2\delta$, the welfare optimal choice and the choice under frame f_L is x', while frame f_S induces the wrong choice x. When $\epsilon_2\delta < \epsilon_1 < \epsilon_2\delta\gamma$, the welfare optimal choice and the choice under frame f_S is x, while frame f_L induces the wrong choice x'.

If the choice problem is framed as a problem of delaying immediate rewards, behavior often reveals greater impatience than if it is framed as a problem of speeding up future rewards. We will use delay framing as an instance of f_S and speed-up framing as an instance of f_L. The literature has proposed different *positive* models to predict the asymmetry between delay and speed-up framing, including the added compensation hypothesis (Benzion et al., 1989), reference-dependence and gain-loss asymmetry (Loewenstein, 1988; Shelley, 1993), or query theory (Weber et al., 2007). Our *normative* analysis only assumes that the delay frame generates present-biased and the speed-up frame generates future-biased choices with respect to welfare.[30]

We conducted an experiment on Amazon Mechanical Turk to obtain individual-level data for a large number of diverse subjects.[31] The experiment took place in August 2016. Our design follows the earlier literature closely; the instructions can be found in Appendix X. After reporting demographic information (gender, age, education), each participant had to answer two pairs of questions about intertemporal choice. Each pair of questions implemented a different frame. In the short-run frame, subjects were asked about their willingness to pay v_0 for an Amazon gift card of given value to be received on the same day. Then they were asked about the minimal discount v_D for which they would accept delaying receipt of the gift card by one year. The answers to this pair of questions reveal the behavioral discount factor $\delta_S = (v_0 - v_D)/v_0$. In the long-run frame, subjects were asked about their willingness to pay v_1 for an Amazon gift card of given value to be received in one year. Then they were asked about the maximal additional fee v_F for which they would accept speeding up receipt of the gift card to the same day. The answers to this pair of questions reveal the behavioral discount factor $\delta_L = v_1/(v_1 + v_F)$. The value of the gift card was \$75 in one frame and \$85 in the other frame, but the assignment of values and the order of the questions were randomized.[32] Responses were not incentivized, but the subjects obtained a compensation of \$0.75 for participation.[33]

Overall, 1059 subjects completed our survey. We dropped 218 subjects who did not obey our instructions or who responded in a way inconsistent with the model.[34] This leaves us with 841

[30] In fact, query theory (Weber et al., 2007), which is prominent in psychology, rests on an explicit description of the internal decision-making process, exactly as required for our normative analysis. In our setting, it postulates that there are reasons for early consumption and reasons for late consumption, which we could capture by a welfare utility function of the form $u(x_1, x_2) = r_1 x_1 + r_2 x_2$. When facing a choice, the decision-maker has to access these reasons from memory. Access happens serially, and the frame determines whether information about the short-run or the long-run is accessed first. Finally, due to "output interference" access is "less successful for later queries than for earlier queries" (p. 517). Denoting by $0 < s \leq 1$ the relative share of reasons retrieved in the second query, we would obtain the behavioral utility functions $u_S(x_1, x_2) = r_1 x_1 + s r_2 x_2$ and $u_L(x_1, x_2) = s r_1 x_1 + r_2 x_2$. This corresponds exactly to our model with $\delta = r_2/r_1$ and $\gamma = 1/s$.

[31] The experiments in Loewenstein (1988), Benzion et al. (1989), Shelley (1993), and Weber et al. (2007) feature between 66 and 208 subjects but individual-level results are not reported.

[32] We chose different values of the gift card in the two frames to avoid suggesting that there was an objectively correct answer to our questions, thereby generating a demand effect for consistency. We still chose the values to be similar to each other because the earlier literature has documented an effect of the stake size on discount rates (e.g. Benzion et al., 1989; Shelley, 1993).

[33] The experiments by Benzion et al. (1989) and Shelley (1993) were not incentivized either. Loewenstein (1988) reports on three different experiments, one of which had real monetary incentives. All three experiments by Weber et al. (2007) were incentivized. A significant framing effect is found in all these studies, and, as we will argue below, our quantitative results are also well in the range of their findings. More generally, Cohen et al. (2016) review the literature on the measurement of time preferences and conclude that there are no significant differences between the results of experiments with and without monetary incentives (p. 32f). We add that the primary goal of our experiment is to illustrate the applicability of our theoretical approach, and one may want to replicate the results in an incentivized experiment to increase the confidence in our findings.

[34] From the beginning, we restricted participation eligibility to US subjects with an experience of at least 500

independent observations. About half of the subjects (44.83%) are female. Ages range between 18 and 77, with a mean of 35 and a median of 32. Subjects' educational backgrounds are diverse, including high school (32.46%), undergraduate degree (48.63%), and graduate degree (18.43%) as the highest completed level of education.

The average of the discount factors δ_S revealed by the subjects in the delay frame is 0.56 ($s = 0.009$). The average of the discount factors δ_L revealed in the speed-up frame is 0.67 ($s = 0.008$). A t-test clearly rejects the null hypothesis that these averages are identical ($p = 0.000$, one-sided). Hence we replicate the earlier finding that average impatience is greater in the delay frame than in the speed-up frame. Our results are also quantitatively within the range of the previous findings.[35] Figure 2.2 is a scatterplot of the individual subjects' behavioral discount factors. The share of fully rational subjects (for whom $\delta_S = \delta_L$) is 6.30%. About one quarter of the subjects (26.28%) exhibit a framing effect opposite to the one conjectured above ($\delta_S > \delta_L$). The correlation between δ_S and δ_L is positive ($\rho = 0.41$) and significant ($p = 0.000$). Column (1) in Table 2.2 reports a linear regression of δ_S on the demographic variables, and column (2) reports the analogous regression with δ_L as the dependent variable. The regressions show that only education has a significant effect on behavioral discount rates, with higher levels of education implying weakly higher discount factors and thus more patient behavior under both frames.[36]

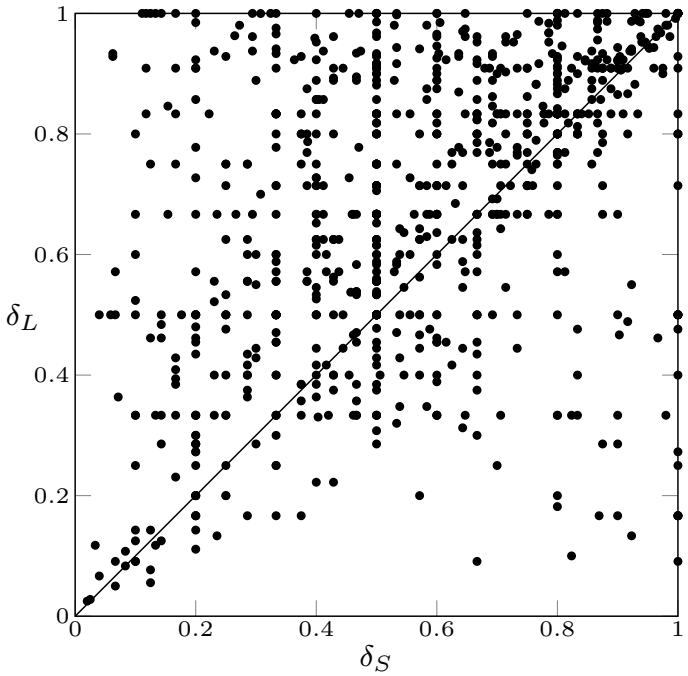
We next examine the welfare discount factors δ and the susceptibility to framing parameters γ that can be deduced from the subjects' behavioral discount rates. The average of δ across subjects is 0.60 ($s = 0.007$), which means that \$1.00 in one year is worth \$0.60 today from an average welfare perspective. Panel (a) of Figure 2.3 shows the entire distribution of δ and reveals considerable heterogeneity across subjects. The average of γ across subjects is 1.17 ($s = 0.013$), and panel (b) of Figure 2.3 shows its distribution in the subject population. The correlation between δ and the framing bias, which we define as $|\gamma - 1|$ to take account of subjects with opposite framing effect, is negative ($\rho = -0.45$) and significant ($p = 0.000$), indicating that the subjects who are more susceptible to framing are those who are less patient from a welfare perspective. Regression (3) in Table 2.2 shows that (only) education affects the welfare discount

MTurk HITs and an approval rate of at least 95%. In the short-run frame, the subjects were instructed to report a discount v_D between 0 and the value v_0 that they had stated earlier. One hundred subjects did not obey this instruction but reported values such that $v_D > v_0$. We deliberately allowed for such responses as a test of (in)attentiveness, following a suggestion by Paolacci, Chandler, and Ipeirotis (2010) for experiments on MTurk. Furthermore, 118 subjects responded in a way that implies either $\delta_S = 0$ or $\delta_L = 0$, which is ruled out in our model.

[35] We report here the average one-year discount factors (δ_S, δ_L) obtained in the previous studies, for the respective treatments that are most similar to ours. Applying our formulas to the average responses in the VCR treatment of Loewenstein (1988) yields $(0.54, 0.80)$. The discount rates reported in Benzion et al. (1989) for the treatment with a \$40 receipt and a one-year time horizon can be translated into the discount factors $(0.72, 0.80)$. Similarly, the pooled results in Shelley (1993) for receipts of \$40 and \$200 and time horizons of 6 months and one year translate into the one-year discount factors $(0.78, 0.83)$. While these values from Benzion et al. (1989) and Shelley (1993) are systematically larger than ours, those from Weber et al. (2007) are lower: the average one-year discount factors in Experiment 1 for gift certificates of values \$50 and \$75 and a time horizon of 3 months are $(0.34, 0.57)$.

[36] The dummy variables "High school", "Undergraduate", and "Graduate" code the highest level of education that a subject has completed. The coefficient "Undergraduate" is significantly larger than the coefficient "High school" in both regressions (1) and (2), but only at marginal significance level in the former (Wald-test, (1) $p = 0.085$, (2) $p = 0.000$). The coefficient "Graduate" is not significantly different from the coefficient "Undergraduate" in both regressions (Wald-test, (1) $p = 0.190$, (2) $p = 0.242$). Hence the effect of education on behavioral patience is only weakly monotonic.

Figure 2.2: Revealed Behavioral Discount Factors



factor, in the expected direction.[37] Maybe surprisingly, the framing bias $|\gamma-1|$ is not significantly affected by education, as can be seen from column (4) in Table 2.2. Hence the effect of education on behavior operates by changing the welfare preference rather than the behavioral bias. The bias is significantly affected by gender, with women exhibiting larger biases, and by age, with smaller biases for older subjects.

We can now address the question of optimal nudging, by combining the theoretical result in Proposition 2.15 with our empirical findings. For transparency, we restrict attention to those subjects for whom we estimated a framing effect $\gamma \geq 1$, but the analysis could easily be extended to the entire subject population. Figure 2.4 shows how many of the subjects should be nudged by one of the two frames, for six different intervals of potential market interest rates. Maybe contrary to conventional wisdom, the short-run frame f_S is an optimal nudge for a substantial share of the subjects across all interest rate conditions (between 74.0% and 95.2%). The share of subjects for whom f_L is optimal is very small (never exceeding 3.5%) and the share of non-nudgeable subjects is limited (between 3.4% and 24.5%). These conclusions are driven by the fact that welfare discount factors are generally low. Recall also that the framing effect is weaker for subjects with greater welfare patience, so framing naturally affects impatient agents more. This can, for instance, be seen in the number of subjects whose behavior would respond to a change in frame for all interest rates in the relevant range. Among those who should be nudged by f_L, the share of such subjects is zero in all the six cases illustrated in Figure 2.4. By contrast, it varies between 0% and 8.1% among the subjects who should be nudged by f_S.

If the regulator's goal is to select a frame that is optimal for a majority of the population, our analysis gives rise to a clear recommendation: choose the frame that induces present-biased

[37] The coefficient "Undergraduate" is significantly larger than the coefficient "High school" (Wald-test, $p = 0.001$), while "Graduate" is not significantly different from "Undergraduate" (Wald-test, $p = 0.890$).

Table 2.2: Regression Analysis

	(1)	(2)	(3)	(4)
Dependent variable	δ_S	δ_L	δ	bias
Constant	0.463***	0.466***	0.443***	0.293***
	(0.051)	(0.074)	(0.046)	(0.078)
Female	-0.002	-0.134	-0.016	0.052**
	(0.017)	(0.017)	(0.015)	(0.025)
Age	-0.001	-0.001	-0.001	−0.002**
	(0.001)	(0.001)	(0.001)	(0.001)
High school	0.077**	0.222***	0.141***	0.054
	(0.036)	(0.065)	(0.032)	(0.063)
Undergraduate	0.111***	0.300***	0.201***	0.022
	(0.035)	(0.064)	(0.031)	(0.062)
Graduate	0.141***	0.274***	0.204***	0.007
	(0.039)	(0.066)	(0.034)	(0.064)
Other controls	Yes	Yes	Yes	Yes
R-squared	0.039	0.036	0.031	0.031
No. of observations	841	841	841	841

Notes: The table reports linear regressions. Robust standard errors are indicated in parentheses. The omitted education category is "None of the others". "Other controls" are response time and dummy variables for the randomization. The symbols *, **, *** denote significance at the 10%, 5% and 1% levels.

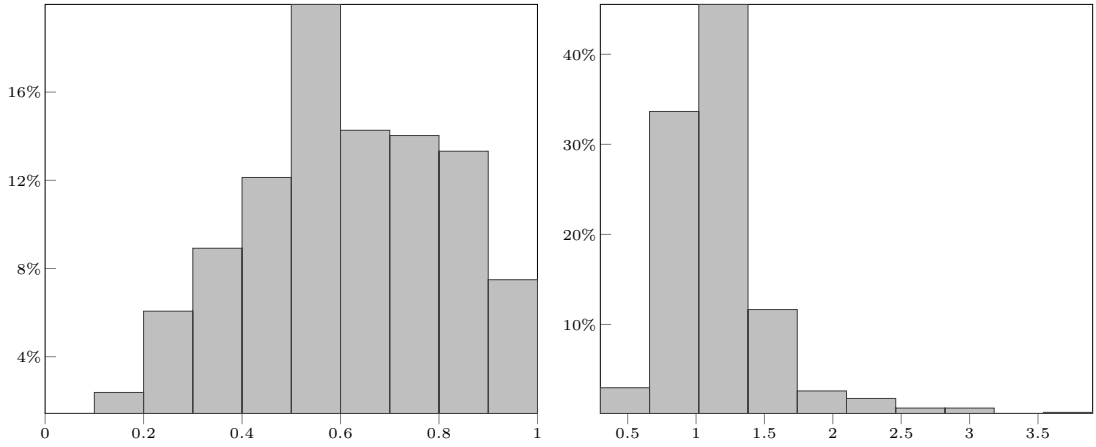
behavior over the one that induces future-biased behavior. A more complete analysis should take into account additional behavioral mechanisms and other consequences of savings decisions (such as externalities for the welfare state), but our results at least challenge the view that soft paternalistic interventions should generally aim at increasing savings.

2.9 Conclusions

We have taken the usual revealed-preference perspective for a single agent. Aside from its methodological justification, this is also directly relevant for nudging, where "personalization does appear to be the wave of the future" (Sunstein, 2014, p. 100). In the digital age of big data, individual-specific data gathering and nudging is achievable, for instance by relying on cookies. However, our results also speak to the problem of nudging a population of agents. On the elicitation stage, an assumption that different agents have identical preferences, possibly after controlling for observables, or are drawn representatively from a population, would allow us to combine observations of different agents into a single data set, facilitating the preference elicitation. On the nudging stage, the necessity to determine one frame for a population of heterogeneous agents gives rise to ordinary social choice problems, which we have refrained from studying in this paper.

Our model-based approach to behavioral welfare economics should in principle be conducive to nudging. Given a conjecture about how agents with different welfare preferences act under

Figure 2.3: Welfare Discount Factors and Susceptibility Parameters



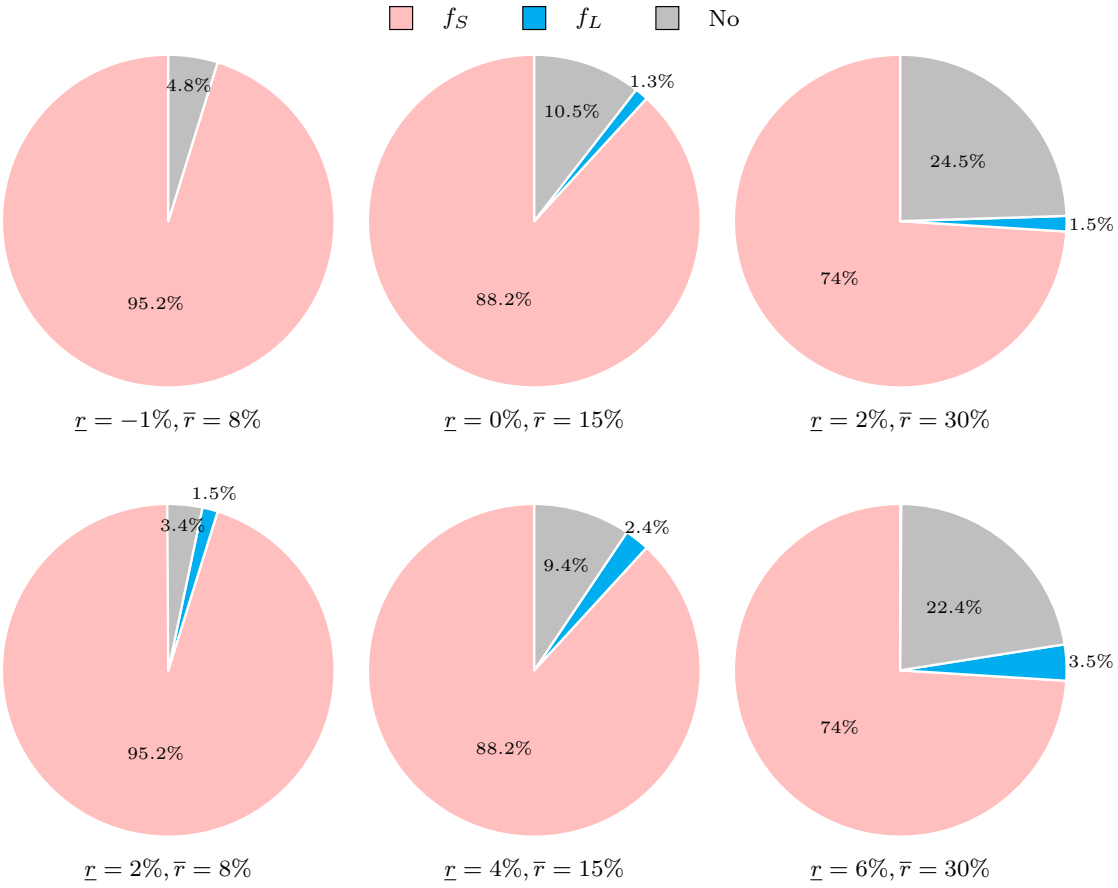
(a) discount factor δ (b) susceptibility γ

different frames, choice data can be used to infer about welfare and to assess which framing of the decision problem helps agents avoid mistakes. It is therefore remarkable how difficult the problem still turns out to be. Welfare-based nudging is impossible for interesting classes of models, and for others it is very complex information-wise. However, our analysis also shows that seemingly small differences between behavioral models can make a big difference for nudging. For instance, a satisficing agent stops searching as soon as some aspiration level is achieved. Our results imply that it is impossible to help this agent make systematically better choices. If the agent stops searching at the end of a search engine's result page, by contrast, it is relatively easy to improve her choices by framing. This raises important questions for future research about actual decision processes.

Acknowledgments

We are grateful for very helpful comments by Sandro Ambühl, Sandeep Baliga, Eddie Dekel, Kfir Eliaz, Samuel Häfner, Igor Letina, Konrad Mierendorff, Georg Nöldeke, Ariel Rubinstein, Yuval Salant, Armin Schmutzler, Ron Siegel, Ran Spiegler, Georg Weizsäcker, seminar audiences at DICE Düsseldorf, European University Institute, Goethe University Frankfurt, HECER Helsinki, Northwestern University, NYU Abu Dhabi, Tel Aviv University, UCL, the Universities of Basel, Bonn, Konstanz, Michigan, Surrey, and Zurich, and participants at CESifo Area Conference on Behavioural Economics 2014, BBE Workshop 2015, Midwest Economic Theory Meeting Fall 2015, Swiss Economists Abroad Meeting 2015, Verein für Socialpolitik Theoretischer Ausschuss 2015, and BERA Micro Workshop 2016. Jean-Michel Benkert would like to thank the UBS International Center of Economics in Society at the University of Zurich and the Swiss National Science Foundation (Doc.Mobility Grant P1ZHP1_161810) for financial support.

Figure 2.4: Optimal Nudging



3 Designing Dynamic Research Contests[1]

Joint with Igor Letina

3.1 Introduction

Research contests[2] have a long history as mechanisms for inducing innovation. From navigation and food preservation, to aviation,[3] research contests have been used to find solutions to some of society's most pressing problems. Recently, the use of research contests by both private and public sector has been expanding rapidly. As in the past, research contests are used in order to foster innovation in some of the most pressing and difficult issues that society is facing. Some examples of problems to which research contests have been applied include vaccine technology, antibiotics overuse, space flight, robotics and AI, as well as environment and energy efficiency.[4] Given that the 2010 America Competes Reauthorization Act authorized US Federal agencies to use prizes and contests, it can be expected that the importance of research contests will only grow in the coming years.

Mistakes in contest design can waste R&D funds and slow the development of important innovations. While some aspects of contest design, like the effect of the number of competitors or the allocation of prizes, are well studied,[5] less is known about the dynamic aspects of contest design. At the same time, as Lang, Seel, and Strack (2014) point out, "there are surprisingly few multi-period contest models in which each player's decision problem is dynamic." This paper deals with the optimal design of dynamic research contests in precisely such an environment. In doing so, we identify a novel design lever that the contest designer can use in order to increase efficiency of the contest – a progress prize, which is paid out in every period until the contest is over and the final prize is awarded.

We are aware of one instance where progress prizes have been used in practice. In 2006, Netflix announced a contest aimed at improving algorithms for making movie recommendations to their users. The contest featured a $1,000,000 Grand Prize which was to be awarded to the first contestant who would improve upon Netflix's own algorithm by at least 10%. More interestingly, a $50,000 Progress Prize could be awarded each year as long as the contest was open.[6] We call this type of contest a progress-prize contest. Despite the wide interest that the

[1] This paper should be cited as Benkert, J.-M. and I. Letina (2017), "Designing Dynamic Research Contests," Mimeo.

[2] Research contests are sometimes referred to as innovation contests or inducement prizes.

[3] 1714 Longitude Prize, 1795 Napoleon's Food Preservation Prize and 1919 Orteig Prize, respectively.

[4] 2012 EU Vaccine Prize; 2015 Better Use of Antibiotics Prize; 1996 Ansari X Prize, 2006 Northrop Grumman Lunar Lander XCHALLENGE and 2007 Google Lunar X-Prize; 2004 DARPA Grand Challenge, 2007 Urban Challenge, 2014 A.I. presented by TED XPRIZE; 1992 Super-Efficient Refrigerator Program, Progressive Insurance Automotive X PRIZE and 2015 NRG Cosia Carbon XPRIZE.

[5] See Che and Gale (2003), Moldovanu and Sela (2001) and Moldovanu and Sela (2006).

[6] See http://www.netflixprize.com.

Netflix Prize attracted,[7] we are (to the best of our knowledge) the first to study progress-prize contests. Moreover, we show that the progress-prize contests are the optimal mechanisms for procuring innovation.

Our setting, which we introduce in Section 3.2, closely follows the seminal work by Taylor (1995). There is a principal who would like to procure an innovation and N agents who can choose to conduct research in each period. Research is costly and yields an innovation whose value is a random draw from some distribution F. The research decisions and the values of innovations are private information to the agents. If an agent reveals the innovation to the principal, the principal can costlessly and accurately determine its value. However, the value is not verifiable by outside parties. In particular, a contract which conditions on the innovation value is not enforceable by courts.[8] In order to incentivize the agents to conduct research, the principal announces a contest. Taylor (1995) studies what we call a fixed-prize contest, in which the principal commits to paying a prize p at the end of an exogenously given deadline of T periods. In a progress-prize contest, the principal commits to both the final prize p (which is paid out once the contest ends), and to the progress prize m (which is paid out to a contestant in each period as long as the contest continues). We endogenize the deadline T, which becomes a design choice of the principal and allow for infinite T. Furthermore, like in the Netflix Prize, the principal can end the contest before the deadline is reached, in which case the final prize p is paid out.

Our first main result (Section 3.3) shows that we can implement any innovation value as the threshold of a global stopping rule with n agents using an appropriate progress-prize contest. In a global stopping rule all n agents conduct research in every period until some agent discovers an innovation with a value above the threshold and then the research effort is stopped by all agents. This result is in sharp contrast to the finding in Taylor (1995), who shows that a fixed-prize contest implements only individual stopping rules. With an individual stopping rule each of the n agents conducts research in every period until she herself discovers an innovation with a value above the threshold, irrespective of what the other agents find. Our result is remarkable for two reasons. First, it eliminates the wasteful duplication of research effort which results with individual stopping rules. Second, contrary to a claim in Taylor (1995, p. 883), we implement a global stopping rule without requiring verifiability of research outcomes.

There is clear economic interpretation underlying our implementation result. On the one hand, the final prize takes care of the agents' incentives. Namely, it induces them to conduct research in every period in order to obtain an innovation of value above the threshold and win the final prize. The presence of the progress prize complicates matters, because an agent with a value above the threshold may wish to delay the submission in the hope of winning the progress prize in this period and then the final prize in the next period. Such a delay is risky, though, as another agent may also obtain an innovation above the threshold and end the contest. A sufficiently high final prize ensures that delaying is not profitable and induces truthful reporting of research outcomes in addition to research effort in all periods. On the other hand, the progress

[7] There were 44,014 valid submissions and the Grand Prize was already awarded less than three years after the contest began. Interestingly, while two of the algorithms which won progress prizes have since been used by Netflix, the winning algorithm was never employed due to too high engineering costs required to implement it. See http://tiny.uzh.ch/F0.

[8] These assumptions are standard in the literature, see for example Taylor (1995) and Che and Gale (2003).

prize takes care of the principal's incentives. In the absence of progress prizes, the principal has no incentive to stop the contest once a value above the threshold has been submitted, because she does not bear the marginal cost of research, but stands to gain by obtaining an even higher value in the next period. The progress prize is set such that it equals the marginal benefit of continuing the contest for one more period when the threshold value has been obtained. Thus, the marginal cost of continuing the contest for one more period, i.e., the progress prize, equals the marginal benefit of continuing the contest for one more period exactly at the threshold. Consequently, the principal wants to stop the contest if and only if a value above the threshold is obtained, thereby giving rise to a global stopping rule.

While the ability to implement global stopping rules using progress-prize contests is of interest in itself, it does not answer the question of whether the principal would actually want to do so. Indeed, one can show that for a given threshold value and number of agents, a global stopping rule has lower expected research costs than an individual stopping rule, but that the latter has a higher expected value of innovation than the former. Hence, it is a priori unclear which contest maximizes the principal's utility. Moreover, it is not clear if some other contest, or some other mechanism performs better than either a progress-prize or a fixed prize contest. Our second main result (Section 3.4) addresses this issue by showing that the principal can implement the first-best outcome using an appropriate progress-prize contest. Thus, progress-prize contests not only outperform fixed-prize contest but constitute the optimal mechanism more generally.

The result builds on ongoing work by Benkert, Letina, and Nöldeke (2017) who consider the problem of finding the optimal search rule in a classic search setting which is very similar to the first-best problem in the present framework. They show that with an infinite search horizon, the searcher will never make use of her ability to recall and thus the solution to the problem coincides with the simpler problem of searching with no recall. In the present context, one can show that the principal optimally chooses an infinite horizon as this comes at no loss and precludes a premature end of the search. Given this, we can make use of the findings in Benkert et al. (2017) and show that the first-best search rule in our context is a global stopping rule with a constant search intensity, which is precisely what we can implement using a progress-prize contest.

In the optimum, the agents continue doing research until an innovation of desired quality is discovered. That is, the principal never imposes a finite deadline on the optimal contest. However, in some applications there may be an exogenous deadline beyond which the contest cannot run or the innovation becomes worthless. We allow for this constraint in Section 3.5 and are once more interested in the performance of a progress-prize contest relative to the first best and to the fixed-prize contest. With a finite deadline the first-best outcome does not have as simple a structure as with an infinite deadline. Gal, Landsberger, and Levykson (1981) and Morgan (1983) have showed that while a global stopping rule is still optimal, the first-best search intensity can generally change non-monotonically over time. Without additional assumptions on the research process it is impossible to say more about the first best. We thus impose a breakthrough innovation structure, which means that we assume that there are essentially two types of innovations – breakthroughs and low-value innovations. This structure represents

a reasonable descriptions for many research contests, as an innovation will often have some dimensions which are key to the principal and make it a breakthrough, while other dimensions affect the value of the innovation in only a negligible fashion. Under the assumption of a breakthrough innovations structure, the first-best search rule is to have all available agents do research until a breakthrough was obtained and to then stop. Hence, the first best is once more a global stopping rule with a constant search intensity. Thus, given the breakthrough innovation structure the progress-prize contest once more not only outperforms the fixed-prize contest, it also implements the first best and thus constitutes the optimal mechanism more generally.

In the absence of the assumption it is not possible to analytically compare the two contests in case of a finite deadline. However, given the finite nature of our setting we can evaluate the two contests numerically. In our simulations the progress-prize contest weakly outperforms the fixed-prize contest in every single instance. Moreover, the progress-prize contest does very well relative to the first-best outcome.

Generally, research contests are classified into fixed-prize contests and innovation races.[9] To win a fixed-prize contest an agent needs to have the best innovation by the end of the contest, whereas an agent needs to have a specific innovation as quickly as possible to win an innovation race. Thus, for an innovation race to be feasible, verifiability is necessary in order to determine whether some proposed innovation is indeed the innovation required to win the race. When innovation races are implemented in practice, a verifiable proxy is commonly used to determine whether an innovation meets the principal's requirements. In the case of the 1996 Ansari X Prize, the objectively verifiable proxy was to have two manned space flights within two weeks using the same spacecraft. The larger objective of the organizer, however, was to "incentivize the creation of a safe, reliable, reusable, privately-financed manned space ship to demonstrate that private space travel is commercially viable".[10] The advantage of a race is that it proceeds until an appropriate innovation has been developed and that it minimizes the wasteful duplication. However, the need to use a proxy may introduce substantial noise. When implementing a fixed-prize contest, the principal in general does not have to use a proxy. However, other problems arise, as the sponsor of a fixed-prize contest needs to announce a deadline. If the agents are not given enough time, they may fail to produce a good enough innovation.[11] If the deadline is very late, however, there is a large risk of wasteful duplication. Our progress-prize contest offers a solution to these problems, as it combines the best of the two formats. It inherits the race-like structure from the global stopping rule, thereby eliminating wasteful duplication and the risk of a premature ending. Moreover, it does so without requiring a noisy proxy. Overall, our results indicate that progress-prize contests could result in substantially more efficient research contests.

[9] See the discussion in Taylor (1995).

[10] See http://ansari.xprize.org.

[11] The objective of the 2004 DARPA Grand Challenge was to "accelerate the development of autonomous vehicle technologies that can be applied to military requirements" but none of the competitors managed to fulfill the requirements of the tournament. Eventually, the requirements were matched in the 2005 DARPA Grand Challenge, suggesting that more time was needed to be successful See the official website on http://archive.darpa.mil/grandchallenge04/.

3.2 Model

3.2.1 Setting

This section introduces the formal framework, which very closely mirrors Taylor (1995). There is a risk-neutral principal who wants to procure an innovation and a finite pool of $N \geq 2$ identical risk-neutral agents who can potentially produce the innovation by conducting research. If the principal obtains the innovation in any period $s \in \{1, \ldots, T\}$ with $T \leq \infty$, her payoff is $\delta^{s-1}\theta - \sum_{t=1}^{T} \delta^{t-1} w_t$, where θ is the value of the innovation and $w_t = \sum_i w_{ti}$ is the sum of transfers made to all agents in period t. Agent i's payoff is $\sum_{t=1}^{T} \delta^{t-1}(w_{ti} - c_{ti})$, where w_{ti} is the transfers received and c_{ti} is the cost incurred through research activities in period t. We assume that the innovations are of no intrinsic value to the agents and allow for any $\delta \in (0, 1]$.

An agent can conduct research in any period t at per-period cost $C > 0$. In each period in which the agent performs research he obtains an innovation of value $\theta \in \Theta$. The innovation value obtained is an independent draw from some distribution F with full and finite support where $\Theta = \{\theta^1, \theta^2, \ldots, \theta^K\}$. Without loss, assume that $\theta^{k+1} > \theta^k$ and normalize $\theta^1 = 0$. Agents can repeatedly conduct research and have perfect recall, that is, they can access all their own previous innovations at any point in time. Initially, every agent is endowed with a worthless innovation and in each period an agent does not conduct research he receives a worthless innovation.

The agents' research activity (whether or not they conduct research in any given period) and research outcomes (the value of an innovation obtained in any given period) are private information. If an agent submits an innovation to the principal, the principal can determine the value of the innovation at no cost. However, the value of an innovation is not verifiable by a court. Thus, contracts conditioning on the value of innovation are not credible.[12]

As noted, this setting very closely mirrors Taylor (1995). We depart from his model in three ways. First, we allow for a possibly infinite horizon, whereas Taylor (1995) considers only the case of an exogenously given finite horizon, which we deal with in Section 3.5. Second, Taylor (1995) only considers the special case of no discounting, while we allow for any $\delta \in (0, 1]$. Third, we assume that the values of innovations are drawn from a discrete distribution F, while Taylor (1995) assumes that the values of innovations are drawn from a continuous distribution with full support. While our main results in Sections 3.3 and 3.4 also hold under the assumption of a continuous distribution, we make use of the discreteness assumption in Section 3.5. To simplify exposition we therefore maintain the discreteness assumption throughout the paper.

3.2.2 Contests

A *progress-prize contest* (PPC) is a tuple $\Gamma = \langle E, p, m, n, T \rangle$, where $E \in \mathbb{R}$ is an entry fee, $p \in \mathbb{R}$ is a final prize, $m \in \mathbb{R}$ is a progress prize, $n \in \mathbb{N}_0$ is the (maximal) number of agents, and $T \in \mathbb{N} \cup \{\infty\}$ is the contest deadline.[13] More specifically, the principal announces the entry fee, the final prize, the progress prize, the number of agents, and the deadline before the

[12] Non-observability and non-verifiability is a typical feature of research activity. As Taylor (1995, p. 873) notes "research inputs are notoriously difficult to monitor" and "courts seldom possess the ability or expertise necessary to evaluate technical research projects".

[13] A deadline of $T = \infty$ should be interpreted as there being no deadline.

contest starts. She then charges the entry fee (or subsidy) at the beginning of the contest from each of the participating agents.[14] Once the contest has started, agents can conduct research and submit innovations to the principal in each period and the principal can decide to end the contest in any period $t \leq T$, but has to end it by period T. If the principal ends the contest in period t, she has to pay out the prize p to one of the participating agents. If the contest does not end in period t, the principal has to pay out the progress prize m to one of the participating agents.

While a progress-prize contest bears some resemblance to the contest in Taylor (1995), there are several major differences. The key departure is the introduction of progress prizes. Further, in Taylor (1995) the deadline T is finite, exogenously given and the principal cannot end the contest before the deadline T. In contrast, the choice of T is endogenous in our model and part of the contest design. In order to implement a PPC contest Γ courts need to be able to verify that the progress prize was paid in every period until the final prize is paid, which can happen no later than at the deadline T. In contrast, the contest in Taylor (1995) only requires that courts can verify whether the final prize was paid out at the deadline. We view this strengthening as uncontroversial, in particular given that the Netflix Prize discussed in the introduction was essentially a PPC. To avoid confusion, we will refer to Taylor's (1995) contest as a *fixed-prize contest* (FPC).

A PPC induces a T-period dynamic game of incomplete information with the set of players being the principal and the agents who participate in contest. Thus, the principal not only announces the contest, but once the contest has started, she is a player in the induced game. The set of players, their payoff functions, the research technology and the contest structure are common knowledge. The agent's research activity and outcomes are private information.

3.3 Implementation of Global Stopping Rules

Taylor (1995) shows that any FPC uniquely implements an *individual stopping rule*, where each agent does research in every period of the contest until an individual threshold value of innovation is reached, irrespective of the innovations discovered by other agents. Such an individual stopping rule consequently entails a risk of duplication of research effort across agents. This type of wasteful duplication could be avoided if a *global stopping rule* was implemented, that is, if all agents stopped doing research once a single agent reaches a certain threshold value of innovation. In general, such a global stopping rule cannot be implemented with an FPC. The reason, as Taylor notes, is that the principal cannot credibly commit to stop the contest after the threshold value has been achieved because she does not bear the marginal cost of continued research, while she stands to benefit from any marginal increase in the value of innovation. Taylor (1995, p. 883) thus concludes that a contest which implements a global stopping rule "always depends on the verifiability of research outcomes". In contrast, our next proposition shows that it is in fact possible to implement any value of innovation as the threshold of a global stopping rule using an appropriate progress-prize contest.

[14] If more than n agent wish to participate, n are selected randomly.

Proposition 3.1 *Any value θ^g can be implemented as the threshold of a global stopping rule with any $n \in \{2, \ldots, N\}$ agents and for any $T \leq \infty$ by using a PPC with sufficiently high final prize p and progress prize $m = p(1-\delta) + \delta\Delta(\theta^g, n) - \theta^g$, where*

$$\Delta(\theta^g, n) = F(\theta^g)^n \theta^g + \sum_{j=g+1}^{K} \left(F(\theta^j)^n - F(\theta^{j-1})^n \right) \theta^j.$$

The proposition shows that for any $\theta^g \in \Theta$, there exists a PPC $\Gamma = \langle E, p, m, n, T \rangle$ such that in a perfect Bayesian equilibrium n agents enter the contest, each agent performs research until she obtains an innovation of value θ^g, and, once this occurs, the agent reports the discovered innovation to the principal, who immediately stops the contest and thus any further research effort, declares a winner and pays out the final prize.[15] Moreover, as long as the deadline is not reached and the principal does not end the contest, a progress prize m is paid out to some agent in every period.[16] The entry fee, which may turn out to be an entry subsidy, ensures that the participation constraint is satisfied so that all n agents want to enter the contest.

The progress prize corresponds exactly to the principal's marginal benefit of continuing the contest one more period when an innovation of value θ^g has been submitted. Thus, through the progress prize the principal incurs a constant marginal cost of one more round of research by the agents. Since the marginal benefit of research to the principal is decreasing in θ, the principal strictly prefers to continue the contest whenever the highest innovation value is below θ^g, strictly prefers to stop it whenever it is above, and is indifferent exactly at the threshold value. As a consequence, the principal will credibly stop the contest if and only if at least the threshold value θ^g was reached, as this value of innovation equalizes the marginal cost and benefit of research to the principal.

As discussed, the progress prize provides the incentives to the principal which are necessary to implement a global stopping rule. Conversely, the final prize gives the incentives to the agents to perform research in every period and to report their research outcomes truthfully. Intuitively, as in the FPC, each agent pursues an individual stopping rule which is determined by the expected prize. Increasing the expected prize increases the individual stopping threshold. If the individual stopping threshold is above the global stopping threshold, the agents will conduct research as long as the contest is ongoing. Similarly, the final prize induces the agents to truthfully report their research outcomes. By not reporting an innovation above the threshold, an agent could win the progress prize in the current period in addition to the final prize in the next period. However, not reporting exposes the agent to the risk that another agent will win in the current period and end the contest. As long as the size of the final prize is sufficiently large relative to the progress prize, the agent will report truthfully. Thus, incentives of the agents can be satisfied

[15] In case of multiple breakthroughs of equal value being reported simultaneously, the principal randomly declares a winner among these innovations.

[16] In the Netflix Prize contest, the annual progress prize was awarded to the agent with the best value of innovation in the current period. In contrast, in the PPC the progress prize is awarded randomly. The reason for this is that the progress prize serves to incentivize the principal to stop the contest as soon as the threshold value has been discovered. The random allocation of the prize simplifies notation and the strategic considerations of the agents. The logic of our argument would hold also if the progress prize was awarded to the agent with the best intermediate value. However, this would require that the principal examines the submissions in every period, something that would be inefficient if there were positive costs associated with evaluation of the value of innovation.

by making the final prize large enough. Since the progress prize is a function of the final prize, the incentives of both the principal and agent can be satisfied simultaneously, such that a global stopping rule results in equilibrium. Finally, the entry fee E can be chosen such that n agents indeed want to participate in such a contest by making it an entry subsidy if necessary.

The above intuition also serves as a sketch of the proof. The first step is to show that the principal does not want to deviate from the equilibrium strategy given the progress prize and the second step is to show that for a sufficiently high final prize the agents do not want to deviate either.

We want to emphasize that the implementation result in Proposition 3.1 does not rely on any assumptions on the research process and lengths of deadlines. As we noted earlier, it also extends to the case of continuous distributions of innovation. However, as is typical in contests, there need to be at least two agents participating in the contest to induce any incentives to exert effort. Interestingly, we could allow for a changing number of agents over time. In practice we observe that the number of participants in a contest may change over time. Typically, participants are being eliminated as the end of the contest draws closer.[17] It is not difficult to extend the result to allow for an arbitrary sequence of agents which is fixed at the start of the contest.

3.4 The First-Best Outcome

Our first main result above shows that, contrary to the findings in Taylor (1995), it is possible to implement global stopping rules without relying on verifiability of research outcomes using a PPC. While this is of interest in itself, the fact that the principal can implement a global stopping rule in principle does not imply that she actually wants to do so. One may suspect that a global stopping rule is preferable to an individual stopping rule, but this need not be true in general. Indeed, when the threshold value and the number of agents are equal, one can show that the individual stopping rule yields a higher expected value of innovation than the global stopping rule. It is thus not obvious if the PPC performs better than the FPC. Moreover, it is not clear if there exists some other contest, or indeed some other mechanism, which performs better than either the PPC or the FPC. To shed light on this, we turn to the question of optimality in this section.

In the absence of informational barriers our framework corresponds to classic search problems (Gal et al., 1981; Morgan, 1983). In these models the searcher can draw multiple samples from the distribution of values in each period taking into account the best value she already has. The only difference is that in our framework there is a given pool of agents N, which acts as a constraint on the number of samples that can be taken in any period. In the absence of such a constraint, Benkert et al. (2017) show that the searcher optimally searches with a constant sample size n^{FB} across time until a constant threshold θ^{FB} has been reached. Put differently, a global stopping rule with constant search intensity is an optimal search strategy. It is not difficult to show that imposing a constraint on the number of samples that can be drawn in a single period does not change the structure of the optimal search strategy. This leads to our

[17] For example, the 2015 NRG COSIA Carbon XPRIZE consists of three rounds. Only up to 15 participants will proceed to the second round and only up to 5 five will proceed to the third round. See http://carbon.xprize.org/about/overview.

second main result.

Proposition 3.2 *If $n^{FB} \geq 2$ the first-best outcome can be implemented using a progress-prize contest.*

The result follows in three steps. First, we show that the result in Benkert et al. (2017) extends to our framework with a finite pool of agents N. Second, in the first-best setting without informational barriers the principal would choose an infinite search horizon, as this avoids the risk of a premature ending of the search process. Given these two observations it follows directly that a PPC with no deadline can implement the first-best search rule which consists of a global stopping rule with a constant number of agents. Third, choosing the entry fee E such that the agents' participation constraint is binding allows the principal to fully extract the first-best surplus. Thus, the PPC not only outperforms the FPC, rather, it is the optimal mechanism more generally.

We noted in the previous section that we need at least two agents participating in the contest to induce any effort. As consequence, we can implement the first best using a PPC only when $n^{FB} \geq 2$. However, a slight twist to the PPC allows us to implement the first-best outcome when $n^{FB} = 1$. More precisely, to implement a global stopping rule with only one agent doing research the principal announces a PPC between a "real agent" and a "fictitious agent". The progress prizes are always paid to the fictitious agent. However, only the real agent can receive the final prize, which occurs only in the case he submits an innovation value above the threshold. This way the progress prize still ensures that the principal will adhere to the global stopping rule, the agent will exert effort in every period in order to obtain the final prize by reaching at least the threshold value, and the fictitious agent has no incentive to exert any research effort.

Taylor (1995) notes that the first best could be achieved if the principal, instead of holding one multi-period contest, held a series of one-period contests. However, if inspecting the agents' submissions is costly, running a sequence of one-period contests and inspecting submissions after every period may be prohibitively costly. This points to another advantage of the PPC. Namely, the principal only has to inspect submissions once and only from the agents who have developed an innovation of high enough quality.

3.5 Exogenous Deadlines

In the last section we have established that we can implement the first-best outcome with a PPC. As we have seen, the result relies on the principal's ability to choose the deadline, that is, to increase the maximal length of the contest. In this section we explore the possibility of an exogenously given finite deadline as is assumed in Taylor (1995), to take into account that there may be situations in which the principal does not have the ability to choose the deadline.

Our implementation result in Proposition 3.1 does not hinge on the principal's ability to choose the deadline. Thus, we can still implement any value of innovation as the threshold of a global stopping rule when the deadline is exogenously given. As we noted, Benkert, Letina and Nöldeke (2017) show that the first-best outcome takes a particularly simple form in the case of an infinite horizon. With an exogenously given finite deadline the first best takes a more complicated form. In general, the first best is characterized by a function $n(\theta, t)$, which

specifies the number of agents which optimally do research as a function of time and the current highest value θ. Gal et al. (1981) and Morgan (1983) have showed that there exists a global stopping value θ^g such that $n(\theta, t) = 0$ for $\theta \geq \theta^g$ and that the number $n(\theta, t)$ is decreasing in θ and increasing in t. In particular, the optimal number of agents doing research will generally change non-monotonically over time.[18] Without additional assumptions on the distribution of innovation qualities it is impossible to say more about the first-best research process. We thus make the following assumption.

Assumption 3.1 (breakthrough innovation structure) *The set of values of innovation is given by* $\Theta = \{0, \varepsilon, 2\varepsilon, \ldots, k\varepsilon, \theta^b, \theta^b + \varepsilon, \theta^b + 2\varepsilon, \ldots, \theta^b + r\varepsilon\}$ *for some* $\theta^b, \varepsilon > 0$ *large and small enough, respectively.*

Given this assumption, any innovation falls into one of two categories – breakthroughs and low-value innovations. The assumption captures the idea that when a principal wants to procure an innovation, there are some key characteristics which largely determine the value of an innovation, while other factors may influence the value but not in a substantial way. For instance, in the Ansari X Prize mentioned in the introduction, the sponsors used the objectively verifiable goal of "build[ing] and launch[ing] a spacecraft capable of carrying three people to 100 kilometers above the Earth's surface, twice within two weeks". While a spacecraft that met this goal could be better or worse, the difference to the organizer did probably not matter as much as achieving that publicly stated goal. Thus, we believe that the breakthrough innovation structure is a good approximation for many research contests.

Given Assumption 3.1, the first best takes a very simple and intuitive form. Essentially, only innovations in which a breakthrough has been achieved are valuable to the principal and all of these have approximately the same worth to the principal. The principal thus wants high research effort as long as no breakthrough was obtained, but no more effort following any breakthrough innovation.

Proposition 3.3 *Given Assumption 3.1,* $n(\theta, t) = N$ *for* $\theta < \theta^b$ *and* $n(\theta, t) = 0$ *for* $\theta \geq \theta^b$.

Thus, given a breakthrough innovation structure, the first best is a global stopping rule in which research effort is maximal until a breakthrough obtains. Proposition 3.3 also obtains under a weaker form of a breakthrough innovation structure than Assumption 3.1, which is given in the proof of Proposition 3.3.

As we noted in Section 3.3, an FPC induces an individual stopping rule. Hence, with a fixed prize being paid out at the end of the contest it is impossible to implement the first best even in the presence of a breakthrough innovation structure, as each agent would try to obtain a breakthrough on their own. In contrast, a PPC can implement any innovation value level as the threshold of a global stopping rule, yielding the following result.

[18] Given the finite time horizon there is the basic trade-off between increasing the chance of getting a high value of innovation by having many agents do research in a given period and risking wasteful duplication. As the deadline approaches the principal becomes more willing to risk duplication for a given value as there are less research opportunities in the future. Having a relatively high value of innovation early on reduces the pressure to get a better innovation in the future and the principal is therefore less willing to risk duplication by having many agents do research simultaneously.

Proposition 3.4 *Given Assumption 3.1, a PPC is optimal and it implements the first best.*

The proof of this result is straightforward. We know from Proposition 3.3 that under Assumption 3.1 the first best is a global stopping rule in which all N agents conduct research in every period until a single breakthrough obtains. Thus, the first best is characterized by a constant number of agents doing research until a global threshold is reached. However, this is precisely what Proposition 3.1 tells us we can achieve using an appropriate PPC.

The breakthrough innovation structure may seem natural in the context of research, however, it need not apply to all settings of interest. The question of what to do when the assumption is not satisfied remains. In general, it is not possible to analytically derive the optimal FPC and PPC, as one would need to solve for the optimal number of agents and the optimal threshold. However, the finite nature of the setting offers the possibility of numerically comparing the two contests. To do so, we ran extensive simulations allowing for a wide range of parameters.[19] More precisely, we randomly drew the size of the support Θ, the values in the support, the distribution of the values, the cost parameter and the discount factor. We consider deadlines ranging from 2 to 20 periods. We find that for every draw of these parameters the PPC weakly outperforms the FPC with an average is 4.9% over all 2000 draws and a maximum of 231%.[20] Turning to the comparison of the PPC with the first-best outcome we find that the PPC does surprisingly well even for short deadlines and achieves about 99.9 % of the expected utility in the first-best outcome.

The results of our simulations clearly demonstrate the superiority of the PPC over the FPC. Moreover, the PPC does remarkably well relative to the first best even with short deadlines. We would expect from our result in Proposition 3.2 that the PPC's performance only improves for longer deadlines as it implements the first best in the absence of a deadline. Indeed, our simulations show that for deadlines longer than 10 periods the PPC essentially matches the first-best outcome in our simulations. Our final result shows that this finding does not extend to the FPC.

Proposition 3.5 *A principal who implements a PPC is strictly better off with later deadlines, while a principal who implements an FPC may be worse off with later deadlines.*

There are two opposing effects at play when the length of the contest is increased. First, the principal benefits because it increases the expected value of the innovation she will eventually obtain. Second, it increases the risk of wasteful research in the form of duplication. The beneficial effect is clearly present in both a PPC and an FPC. However, in contrast to the FPC, there is no change in the amount of duplication when a global stopping rule is implemented using a PPC, as it ensures that research stops conditional on reaching the threshold. This effect is not present with an individual stopping rule which results under an FPC. Thus, depending on which effect dominates, increasing the duration of the contest may be harmful for a principal using an FPC, whereas the principal is unambiguously better off using a PPC.

[19] The simulations were conducted in R. The code is available from the authors upon request.

[20] For the comparison between the PPC and the FPC we have to set $\delta = 1$, as the results on the FPC in Taylor (1995) do not extend to the case with discounting. In the case of discounting the difference between the two contests would only be magnified, as the FPC cannot be ended early.

3.6 Related Literature

The seminal paper on dynamic research contests is Taylor (1995) on which we build our model.[21] He shows that a T-period FPC with N agents uniquely implements an individual stopping rule among the agents. Further, Taylor shows that it is optimal to limit the number of agents in the contest and that the principal can extract the entire ex ante surplus using appropriate entry fees. Fullerton, Linster, McKee, and Slate (2002) compare Taylor's FPC to an auction in the same setting. That is, the only change to Taylor's framework is that at the end of the contest an auction is used to allocate the prize among the agents. The authors argue that this lowers the principal's informational requirements and they provide experimental evidence that this is more cost-effective when the principal cannot charge entry fees. Rieck (2010) considers a variation of Taylor's framework which enables him to study the role of information revelation. He shows that when the agents' research outcomes are publicly revealed there are essentially two stopping thresholds instead of one. If the highest quality among the agents is above the upper threshold all agents stop research, if the highest quality is between the two thresholds only the leading agent stops research and if all qualities are below the lower threshold all agents continue to do research. Depending on the parameters the principal may be better off with or without information revelation.

Recently, a number of papers have used bandit models to study the problem of incentive provision for dynamic research activity. In bandit models it is unclear ex ante if the innovation in question can actually be successfully realized, so in contrast to our setting these models focus on learning over time. Halac, Kartik, and Liu (forthcoming) consider the optimal design of contests for innovation where the principal chooses the prize-sharing scheme and a disclosure policy which determines what information is revealed to the agents about their respective outcomes. Similarly to our setting, the first-best features a global stopping rule. However, they find that a contest which does not entail a global stopping rule can be optimal in the presence of learning. In particular, in a broad class of contests it is optimal to stop the contest only once a certain number of agents had a success and to share the prize between them.[22] Along similar lines Green and Taylor (2016) consider the role of breakthroughs in a single-agent contracting environment. In contrast to our framework, the research outcome can be contracted upon and the problem the principal faces is how to optimally induce effort over time using a first deadline for the breakthrough, a second deadline for the final outcome and a monetary transfer. In their paper the monetary transfer is decreasing over time, which induces the agent to aim for an early success. Thus, the slope of the prize schedule is used to affect the agent's incentives. In contrast, in our paper the final prize aligns the agents' incentives, while the progress prizes over time serve to align the principal's incentives.

Related is also the literature on optimal design of research contests in the static setting, where the seminal contribution is Che and Gale (2003). In their paper the contest design consists of the choice of the set from which the agents can choose their prize. They show that with symmetric agents the optimal contest is an auction and the optimal number of agents is two. When agents

[21] Konrad (2009) provides an excellent overview of the literature on contests. See also Siegel (2009) for general results on all-pay auctions.

[22] For a related model featuring partial progress see Bimpikis, Ehsani, and Mostagir (2014).

are asymmetric, the optimal contest is still an auction with two agents, but the optimal auction handicaps the more efficient agents.[23] The major difference between Che and Gale (2003) and our paper is that we focus on the dynamic aspects of contest design. Thus, the question of wasteful duplication of effort over time, which is the central issue we address with the PPC, does not arise in Che and Gale (2003). Another difference is the choice of innovation technology — when innovation is deterministic, as in Che and Gale (2003), there is no sampling benefit from having more than two agents.[24] Additionally, an auction gives market power to the agents, and when the innovation technology is sufficiently random, an auction might perform badly as the agent profits from the good realizations.[25] Several other directions have been explored in the static setting. Letina and Schmutzler (2016) consider the optimal contest design when the agents can choose their approach to innovation and the principal attempts to give them incentives to diversify their approaches because of the resulting option value. They find that the optimal contest is what they call a bonus tournament, where a winner gets a fixed prize, plus a bonus if he outperforms the second best agent with a high enough margin.

Lang et al. (2014) is related to our result about the optimal contest length T. They consider a two-player, FPC where agents exert effort over time and breakthroughs arrive according to a Poisson process. The agent with the most breakthroughs wins. They find that the principal can be better off with a shorter deadline. In our paper, the principal benefits from longer deadlines with a PPC contest and may be better or worse off with an FPC.

Related to our implementation result is Kruse and Strack (2015). They study a dynamic principal-agent problem, where the agent observes realizations of a stochastic process over time. They show that for any threshold value, the principal can induce the agent to stop the game as soon as the process is above the threshold by committing to an appropriate schedule of transfers which depend only on the period when the game is stopped. In our paper, the stochastic process comes from the research done by the agents and the goal of the contest is to incentivize the agents to engage in research.

There is relatively little empirical work on dynamic research contests.[26] Using data on software contests Boudreau, Lacetera, and Lakhani (2011) find that increasing the number of participants reduces average effort but increases the chance of getting a very high quality innovation. Also using data on software contests Boudreau, Lakhani, and Menietti (2016) find that the results derived in Moldovanu and Sela (2001) generally perform quite well. In particular, they find that the response of participants to an increase in the number of competitors yields heterogeneous responses. Namely, low ability agents respond weakly, medium ability agents decrease their efforts while high ability agents increase their efforts.

3.7 Conclusion

The goal of the present paper is to increase our understanding of the optimal design of research contests, which have recently seen a rapid expansion in practice. Research contests are inherently

[23] Discrimination in contests is also studied in Pérez-Castrillo and Wettstein (2016).

[24] See for example Terwiesch and Xu (2008) and Letina (2016).

[25] This is the case in Schöttner (2008) who shows that when innovation technology is sufficiently random, a research tournament can outperform an auction.

[26] For a recent survey of experimental work on contests see Dechenaux, Kovenock, and Sheremeta (2015).

dynamic by the nature of research itself and by virtue of the contest taking place over a longer period of time. It turns out that taking into account the dynamic nature of research improves contest design dramatically. Indeed we show that the introduction of progress prizes yields a substantial improvement for the principal. In particular, such a progress-prize contest constitutes the optimal mechanism and allows the principal to implement the first best if she can freely choose the deadline of the contest. Our results show that even when the deadline is exogenously given, a progress-prize contest does better than the classic fixed-prize contest in the seminal work by Taylor (1995) and can even implement the first best under some conditions on the research process.

Interestingly, an alternative but similar contest which does not entail progress prizes but instead features a dynamic prize schedule does almost as well as the progress-prize contest. In this alternative contest, the principal commits to a sequence of prizes $p_1, ..., p_T$. If the principal ends the contest in period t, the prize p_t has to be paid out. The difference between two prizes p_t and p_{t+1} plays a similar role to our progress prize. This alternative contest with a dynamic prize schedule can also implement global stopping rules, but only with finite deadlines. Thus, the first-best outcome can be approximated arbitrarily well, but not fully achieved.

We believe that our results have important implications for the design of research contests as they promise a substantial improvement over the standard fixed-prize contest. As we already noted in the introduction, a progress-prize contest allows us to combine the best aspects of two commonly employed contest formats, fixed-prize contest and innovation races. In particular, progress prizes should be easy to implement in most contests as the Netflix Prize has demonstrated.

Acknowledgments

We are grateful for very helpful comments by Eddie Dekel, Christian Ewerhart, Andreas Hefti, Botond Kőszegi, David Levine, Shuo Liu, Nick Netzer, Georg Nöldeke, Yuval Salant, Armin Schmutzler, Philipp Strack and seminar audiences at the Swiss IO Day 2016, EARIE 2016, the Barcelona GSE Summer Forum 2016, the VfS 2016, the Swiss Theory Day 2016, and the Zurich Workshop on Economics 2016. Jean-Michel Benkert would like to acknowledge the hospitality of Northwestern University where some of this work was carried out and the Swiss National Science Foundation (Doc.Mobility grant P1ZHP1_161810) as well as the UBS International Center of Economics in Society at the University of Zurich for financial support. Igor Letina would like to acknowledge the hospitality of Stanford University where some of this work was carried out and the Swiss National Science Foundation (Doc.Mobility grant P1ZHP1_155283) for financial support.

Part III

Appendices

A Appendix: Chapter 1

A.1 Proofs

A.1.1 Impossibility Result

We begin by noting that

$$\begin{aligned}
&\tilde{v}_B(\theta_B) \\
&= \int_{a_S}^{b_S} y^f(\theta_S, \theta_B)\, dF_S(\theta_S) + \eta_B^1 \int_{a_S}^{b_S} \int_{a_S}^{b_S} \mu_B^1 \left(y^f(\theta_S, \theta_B) - y^f(\theta'_S, \theta_B) \right)\, dF_S(\theta'_S)\, dF_S(\theta_S), \\
&= y_B(\theta_B) + \eta_B^1 \int_{a_S}^{b_S} \int_{a_S}^{b_S} y^f(\theta_S, \theta_B)(1 - y^f(\theta'_S, \theta_B)) - \lambda_B^1 (1 - y^f(\theta_S, \theta_B)) y^f(\theta'_S, \theta_B)\, dF_S(\theta'_S)\, dF_S(\theta_S), \\
&= y_B(\theta_B)(1 + \Lambda_B(y_B(\theta_B) - 1))
\end{aligned}$$

and analogously $\tilde{v}_S(\theta_S) = y_S(\theta_S)(1 - \Lambda_S(y_S(\theta_S) - 1))$, where

$$y_B(\theta_B) = \int_{a_S}^{b_S} y^f(\theta_S, \theta_B) dF_S(\theta_S), \quad y_S(\theta_S) = \int_{a_B}^{b_B} y^f(\theta_S, \theta_B) dF_B(\theta_B).$$

Imposing CPEIC we can write the sum of the agents' ex ante expected utilities as

$$\begin{aligned}
&\int_{a_B}^{b_B} U_B(\theta_B) f_B(\theta_B) d\theta_B + \int_{a_S}^{b_S} U_S(\theta_S) f_S(\theta_S) d\theta_S \\
&= U_B(a_B) + \int_{a_B}^{b_B} \int_{a_B}^{\theta_B} y_B(t)(1 + \Lambda_B(y_B(t) - 1)) dt f_B(\theta_B) d\theta_B \\
&+ U_S(b_S) + \int_{a_S}^{b_S} \int_{\theta_S}^{b_S} y_S(t)(1 - \Lambda_S(y_S(t) - 1)) dt f_S(\theta_S) d\theta_S \\
&= U_B(a_B) + \int_{a_B}^{b_B} y_B(\theta_B)(1 + \Lambda_B(y_B(\theta_B) - 1))(1 - F_B(\theta_B)) d\theta_B \\
&+ U_S(b_S) + \int_{a_S}^{b_S} y_S(\theta_S)(1 - \Lambda_S(y_S(\theta_S) - 1)) F_S(\theta_S) d\theta_S.
\end{aligned}$$

Note that the monotonicity constraints are satisfied due to Assumption 1, i.e., $\Lambda_B, \Lambda_S \leq 1$. Further, from Lemmas 1.1 and 1.2 and the corresponding discussion in the main text we know that we can set the loss aversion in the money dimension to zero. This allows us to express the sum of the agents' ex ante expected utilities as

$$\begin{aligned}
&\int_{a_B}^{b_B} U_B(\theta_B) f_B(\theta_B) d\theta_B + \int_{a_S}^{b_S} U_S(\theta_S) f_S(\theta_S) d\theta_S \\
&= \int_{a_B}^{b_B} \int_{a_S}^{b_S} (\theta_B - \theta_S) y(\theta_S, \theta_B) f_S(\theta_S) f_B(\theta_B) d\theta_S d\theta_B \\
&+ \int_{a_S}^{b_S} \theta_S y_S(\theta_S) \Lambda_S(y_S(\theta_S) - 1) f_S(\theta_S) d\theta_S + \int_{a_B}^{b_B} \theta_B y_B(\theta_B) \Lambda_B(y_B(\theta_B) - 1) f_B(\theta_B) d\theta_B
\end{aligned}$$

where we used CPEIC and integration by parts towards the end. Putting these two equations together we get

$$
\begin{aligned}
&U_B(a_B) + U_S(b_S) \\
&= \int_{a_B}^{b_B} \int_{a_S}^{b_S} (\theta_B - \theta_S) y(\theta_S, \theta_B) f_S(\theta_S) f_B(\theta_B) d\theta_S d\theta_B \\
&+ \int_{a_S}^{b_S} \theta_S y_S(\theta_S) \Lambda_S(y_S(\theta_S) - 1) f_S(\theta_S) d\theta_S + \int_{a_B}^{b_B} \theta_B y_B(\theta_B) \Lambda_B(y_B(\theta_B) - 1) f_B(\theta_B) d\theta_B \\
&- \int_{a_B}^{b_B} y_B(\theta_B)(1 + \Lambda_B(y_B(\theta_B) - 1))(1 - F_B(\theta_B)) d\theta_B - \int_{a_S}^{b_S} y_S(\theta_S)(1 - \Lambda_S(y_S(\theta_S) - 1)) F_S(\theta_S) d\theta_S.
\end{aligned}
$$

Individual rationality requires $U_B(a_B) + U_S(b_S) \geq 0$. We will now show that this condition is never satisfied for any combination of buyer and seller loss aversion. From our discussion in the main text, we know that it is sufficient to consider the case $\Lambda_S = 0$, i.e., no loss aversion on the trade-dimension for the seller. This allows us to simplify and rewrite to

$$
\begin{aligned}
&U_B(a_B) + U_S(b_S) \\
&= \int_{a_B}^{b_B} \int_{a_S}^{b_S} \left(\left[\theta_B - \frac{1 - F_B(\theta_B)}{f_B(\theta_B)} \right] - \left[\theta_S + \frac{F_S(\theta_S)}{f_S(\theta_S)} \right] \right) y(\theta_S, \theta_B) f_B(\theta_B) f_S(\theta_S) d\theta_S d\theta_B \\
&+ \Lambda_B \int_{a_B}^{b_B} y_B(\theta_B)(y_B(\theta_B) - 1) \left[\theta_B - \frac{1 - F_B(\theta_B)}{f_B(\theta_B)} \right] f_B(\theta_B) d\theta_B.
\end{aligned}
$$

MS show in their proof of Theorem 1 (p. 269) that

$$
\begin{aligned}
&\int_{a_B}^{b_B} \int_{a_S}^{b_S} \left(\left[\theta_B - \frac{1 - F_B(\theta_B)}{f_B(\theta_B)} \right] - \left[\theta_S + \frac{F_S(\theta_S)}{f_S(\theta_S)} \right] \right) y(\theta_S, \theta_B) f_B(\theta_B) f_S(\theta_S) d\theta_S d\theta_B \\
&= - \int_{a_B}^{b_S} (1 - F_B(x)) F_S(x) \, dx.
\end{aligned}
$$

Further, we have $y_B(\theta_B) = F_S(\theta_B)$ since we are considering the ex post efficient mechanism. Putting this together yields

$$
\begin{aligned}
U_B(a_B) + U_S(b_S) = &- \int_{a_B}^{b_S} (1 - F_B(x)) F_S(x) \, dx \\
&+ \Lambda_B \int_{a_B}^{b_B} F_S(x)(F_S(x) - 1) \left[x - \frac{1 - F_B(x)}{f_B(x)} \right] f_B(x) dx.
\end{aligned}
$$

Careful inspection of the limits of the integrals shows that

$$
\begin{aligned}
&U_B(a_B) + U_S(b_S) = - \int_{\max\{a_B, a_S\}}^{\min\{b_S, b_B\}} (1 - F_B(x)) F_S(x) \, dx \\
&+ \Lambda_B \int_{\max\{a_B, a_S\}}^{\min\{b_S, b_B\}} F_S(x)(F_S(x) - 1) \left[x - \frac{1 - F_B(x)}{f_B(x)} \right] f_B(x) dx \\
&= - \int_{\max\{a_B, a_S\}}^{\min\{b_S, b_B\}} (1 - F_B(x)) F_S(x) + \Lambda_B (1 - F_S(x)) F_S(x) \left[x - \frac{1 - F_B(x)}{f_B(x)} \right] f_B(x) dx \\
&= - \int_{\max\{a_B, a_S\}}^{\min\{b_S, b_B\}} (1 - F_B(x)) F_S(x)(1 - \Lambda_B(1 - F_S(x))) + \Lambda_B (1 - F_S(x)) F_S(x) x f_B(x) \, dx \\
&< 0,
\end{aligned}
$$

violating individual rationality. To conclude the proof, recall from our discussion of the information rents, that loss aversion in the money dimension makes the problem unambiguously harder, as it reduces the gains from trade without affecting the information rents. Thus, impossibility in the absence of loss aversion in the money dimension implies impossibility in the presence of loss aversion in the money dimension.

A.1.2 Maximizing the Designer's Revenue

Step 1. We begin by imposing CPEIC. In order for the CPEIC constraint to be satisfied, conditions (i) and (ii) from Proposition 1.2 must be satisfied. Using the utility functions given in equations (1.4) and (1.5) from condition (ii), we can rewrite the objective function in the problem (RM) to

$$\int_{a_B}^{b_B} \left(\eta_B^2 w_B(\theta_B) + \theta_B \tilde{v}_B(\theta_B) - U_B(a_B, s_S^t | a_B) - \int_{a_B}^{\theta_B} \tilde{v}_B(t)\, dt \right) dF_B(\theta_B)$$
$$+ \int_{a_S}^{b_S} \left(\eta_S^2 w_S(\theta_S) - \theta_S \tilde{v}_S(\theta_S) - U_S(b_S, s_B^t | b_S) - \int_{\theta_S}^{b_B} \tilde{v}_S(t)\, dt \right) dF_S(\theta_S).$$

From the IR constraint we have $U_B(a_B, \theta_S | a_B) \geq 0$ and $U_S(b_S, \theta_B | b_S) \geq 0$, which enter the objective function negatively. Since we are maximizing the objective function, we choose transfers such that $U_B(a_B, \theta_S | a_B) = 0$ and $U_S(b_S, \theta_B | b_S) = 0$. If the expected utility of these "worst" types was not equal to zero in the optimal mechanism, we could modify the transfers by adding lump-sum transfers and reduce their expected utility to zero without affecting CPEIC. Moreover, w_B and w_S, which are negative by the arguments in the main text, enter positively. Thus, we impose an additional restriction on transfers, namely that they are interim deterministic, which leads to $w_B(\theta_B) = w_S(\theta_S) = 0$ for all $\theta_B, \theta_S \in [a, b]$. Note that these two restrictions on transfers do not contradict each other. Given this, the problem reduces to

$$\max_{(y^f)} \int_{a_B}^{b_B} \left(\theta_B \tilde{v}_B(\theta_B) - \int_{a_B}^{\theta_B} \tilde{v}_B(t)\, dt \right) dF_B(\theta_B)$$
$$+ \int_{a_S}^{b_S} \left(-\theta_S \tilde{v}_S(\theta_S) - \int_{\theta_S}^{b_S} \tilde{v}_S(t)\, dt \right) dF_S(\theta_S)$$
$$\text{subject to } \tilde{v}_S \text{ being non-increasing, } \tilde{v}_B \text{ being non-decreasing,}$$

which proves Proposition 1.4.

Step 2. We next impose that types are uniformly distributed on $[a, a+1]$ and rewrite the objective function in this reduced problem. Using integration by parts we get

$$\int_a^b \left(\theta_B \tilde{v}_B(\theta_B) - \int_a^{\theta_B} \tilde{v}_B(t)\, dt \right) d\theta_B + \int_a^b \left(-\theta_S \tilde{v}_S(\theta_S) - \int_{\theta_S}^b \tilde{v}_S(t)\, dt \right) d\theta_S$$
$$= \int_a^b (2\theta_B - 1 - a) \tilde{v}_B(\theta_B)\, d\theta_B - \int_a^b (2\theta_S - a) \tilde{v}_S(\theta_S)\, d\theta_S.$$

Further, we can write

$$\begin{aligned}\tilde{v}_B(\theta_B) &= \int_a^b y^f(\theta_S,\theta_B)\,d\theta_S + \eta_B^1 \int_a^b \int_a^b \mu^1\left(y^f(\theta_S,\theta_B) - y^f(\theta'_S,\theta_B)\right)\,d\theta'_S\;\,d\theta_S \\ &= y_B(\theta_B) + \eta_B^1\left[y_B(\theta_B)(1-y_B(\theta_B)) - \lambda_B^1(1-y_B(\theta_B))y_B(\theta_B)\right] \\ &= y_B(\theta_B) + y_B(\theta_B)\Lambda_B\left(y_B(\theta_B)-1\right) \\ &= y_B(\theta_B)(1+\Lambda_B\left(y_B(\theta_B)-1\right),\end{aligned}$$

where $\int_a^b y^f(\theta_S,\theta_B)\,d\theta_S = y_B(\theta_B)$. Analogously, we can write $\tilde{v}_S(\theta_S) = y_S(\theta_S)(1-\Lambda_S(y_S(\theta_S)-1))$. Note that therefore the constraints that $\tilde{v}_S$ is non-increasing and $\tilde{v}_B$ non-decreasing are equivalent to y_S being non-increasing and y_B being non-decreasing given the assumption that gain-loss utiltiy does not dominate. Thus, we have reduced the maximization problem to

$$\begin{aligned}\max_{y^f\in\mathcal{Y}} &\int_a^b (2\theta_B - 1 - a)y_B(\theta_B)(1+\Lambda_B\left(y_B(\theta_B)-1\right)\,d\theta_B \\ &- \int_a^b (2\theta_S - a)y_S(\theta_S)(1-\Lambda_S(y_S(\theta_S)-1))\,d\theta_S, \qquad \text{(RM')} \\ &\text{subject to } y_B \text{ being non-decreasing and } y_S \text{ being non-increasing.}\end{aligned}$$

Step 3. We will make use of the reduced-form approach as in Che et al. (2013) to maximize directly over the interim trade probabilities y_B and y_S instead of the ex post allocation rule y^f. First, we perform a change of variables to rewrite the objective function to

$$\max_{y^f\in\mathcal{Y}} \int_0^1 (2x-1+a)q_B(x)(1+\Lambda_B\left(q_B(x)-1\right)\,dx - \int_0^1 (2x+a)q_S(x)(1-\Lambda_S(q_S(x)-1))\,dx,$$

where $q_i(x) = y_i(x+a)$ for all $x\in[0,1]$. Making use of Corollary 6 in Che et al. (2013), we maximize directly over q_B and q_S subject to an allocation and an aggregate constraint. The problem then reads

$$\max_{q_B,q_S} \int_0^1 (2x-1+a)q_B(x)(1+\Lambda_B\left(q_B(x)-1\right)\,dx - \int_0^1 (2x+a)q_S(x)(1-\Lambda_S(q_S(x)-1))\,dx,$$

subject to q_B being non-decreasing, q_S being non-increasing, the allocation constraint

$$\int_{\theta_S}^1 (1-q_S(t))\,dt + \int_{\theta_B}^1 q_B(t)\,dt \le 1-\theta_B\theta_S$$

for all $(\theta_B,\theta_S)\in[0,1]^2$ and the aggregate constraint

$$\int_0^1 (1-q_S(t))\,dt + \int_0^1 q_B(t)\,dt = 1.$$

The allocation constraint is the condition known from Border (1991) and aggregate constraint ensures that the good is either allocated to the buyer or the seller. Following the proof of Lemma 4 in Mierendorff (2016) we can rewrite the allocation constraint to

$$\int_{\theta_S}^1 (1-q_S(t))\,dt \le \min_{\theta_B\in[0,1]}\left[1-\theta_S\theta_B - \int_{\theta_S}^1 q_B(t)\,dt\right]$$

for all $\theta_B \in [0,1]$ and since we are minimizing a convex function on the right-hand side, we obtain

$$\int_{\theta_S}^{1}(1-q_S(t))\,dt \leq 1-q_B^{-1}(\theta_S)\theta_S - \int_{y_B^{-1}(\theta_S)}^{1} q_B(t)\,dt$$

for all $\theta_S \in [0,1]$. This constraint is satisfied with equality when $q_S^*(t) = 1-q_B^{-1}(t)$, where q_B^{-1} denotes the generalized inverse. In what follows, we will show that for a given, non-decreasing function q_B, the function $q_S^*(t) = 1-q_B^{-1}(t)$ minimizes

$$\int_0^1 (2x+a)q_S(x)(1-\Lambda_S(q_S(x)-1))\,dx$$

subject to the allocation and aggregate constraint and to q_S being non-increasing. This implies that is enough to maximize over the set of all non-decreasing trade probabilities q_B such that $q_S(t) = 1-q_B^{-1}(t)$. Consider some other candidate to the solution, $\tilde{q}_S$ which satisfies the allocation constraints and is different from q_S^* on a set of positive measure. Then there must exist an interval $[\underline{u},\bar{u}]$ such that

$$\int_{\theta_S}^{1}(1-\tilde{q}_S(t))\,dt < \int_{\theta_S}^{1}(1-q_S^*(t))\,dt$$

for all $\theta_S \in [\underline{u},\bar{u}]$. We will now construct a function $\hat{q}_S$ which does better than the candidate $\tilde{q}_S$, thereby proving that q_S^* is indeed optimal. To do this, we show that there exist $\bar{p},\underline{p} \in [0,1]$ and $p \in (\underline{p},\bar{p})$ such that (1) $\hat{q}_S(t) = \tilde{q}_S(t)$ for all $t \notin [\underline{p},\bar{p}]$, (2) $\hat{q}_S(t) \geq \tilde{q}_S(t)$ for all $t \in [p,\bar{p}]$, (3) $\hat{q}_S(t) \leq \tilde{q}_S(t)$ for all $t \in [\underline{p},p)$, (4)

$$\int_{\underline{p}}^{\bar{p}} q_S^*(t)-\tilde{q}_S(t)\,dt = 0,$$

and (5)

$$\int_{\theta_S}^{1}(1-\hat{q}_S(t))\,dt \leq \int_{\theta_S}^{1}(1-q_S^*(t))\,dt$$

for all $\theta_S \in [0,1]$. Fix some $\bar{p} \in [\underline{u},\bar{u}]$ and define $\hat{q}_S(t) = \tilde{q}_S(t)$ for all $t > \bar{p}$. Note that by the aggregate constraint there must exist $0 \leq p < \bar{p}$ such that

$$\int_{p}^{1}(1-\hat{q}_S(t))\,dt = \int_{p}^{1}(1-q_S^*(t))\,dt$$

when $\hat{q}_S(t) = \tilde{q}_S(\bar{p})$ for all $t \in [p,\bar{p}]$. This construction satisfies the monotonicity and the allocation constraint. If there now exists a $0 \leq \underline{p} < p$ such that $\hat{q}_S(t) = \tilde{q}_S(\underline{p})$ for all $t \in [\underline{p},p)$ and $\hat{q}_S(t) = \tilde{q}_S(t)$ for all $t < \underline{p}$ with

$$\int_{\underline{p}}^{\bar{p}} \hat{q}_S(t)-\tilde{q}_S(t)\,dt = 0$$

we are done. If not, then we must have even with $\underline{p} = 0$ that

$$\int_p^{\bar{p}} \hat{q}_S(t) - \tilde{q}_S(t)\ dt + \int_0^p \hat{q}_S(t) - \tilde{q}_S(t)\ dt > 0.$$

If

$$\int_p^{\bar{p}} \hat{q}_S(t) - \tilde{q}_S(t)\ dt - \int_0^p \tilde{q}_S(t)\ dt < 0,$$

then there must exist $c > 0$ such that $\hat{q}_S(t) = c$ for $t \in [0, p)$ yields

$$\int_p^{\bar{p}} \hat{q}_S(t) - \tilde{q}_S(t)\ dt + \int_0^p \hat{q}_S(t) - \tilde{q}_S(t)\ dt = 0.$$

If not, then increase p until

$$\int_p^{\bar{p}} \hat{q}_S(t) - \tilde{q}_S(t)\ dt - \int_0^p \tilde{q}_S(t)\ dt = 0.$$

Such a p exists and the such constructed $\hat{q}_S$ satisfies the above (1) to (5). Thus, we have constructed $\hat{q}_S$ from $\tilde{q}_S$ by shifting trade probability from high types to low types, while satisfying the allocation constraint. This was possible, because $\tilde{q}_S$ is different from q_S^* on a set of positive measure and the aggregate constraint needs to be satisfied.

We will now show, that

$$\int_0^1 (2x+a)\hat{q}_S(x)\left(1 - \Lambda_S\left[\hat{q}_S(x) - 1\right]\right)\ dx \leq \int_0^1 (2x+a)\tilde{q}_S(x)\left(1 - \Lambda_S\left[\tilde{q}_S(x) - 1\right]\right)\ dx,$$

implying that $\tilde{q}_S$ cannot be a minimizer. We have

$$\begin{aligned}
&\int_0^1 (2x+a)\left(\hat{q}_S(x)\left(1 - \Lambda_S\left[\hat{q}_S(x) - 1\right]\right) - \tilde{q}_S(x)\left(1 - \Lambda_S\left[\tilde{q}_S(x) - 1\right]\right)\right)\ dx \\
&= \int_{\underline{p}}^{\bar{p}} (2x+a)\left(\hat{q}_S(x)\left(1 - \Lambda_S\left[\hat{q}_S(x) - 1\right]\right) - \tilde{q}_S(x)\left(1 - \Lambda_S\left[\tilde{q}_S(x) - 1\right]\right)\right)\ dx
\end{aligned}$$

by our construction of $\hat{q}_S$. Furthermore, whenever $\hat{q}_S(x) > \tilde{q}_S(x)$, we also have $\hat{q}_S(x)\left(1 - \Lambda_S\left[\hat{q}_S(x) - 1\right]\right) > \tilde{q}_S(x)\left(1 - \Lambda_S\left[\tilde{q}_S(x) - 1\right]\right)$. Thus, we obtain

$$\begin{aligned}
&\int_{\underline{p}}^{\bar{p}} (2x+a)\left(\hat{q}_S(x)\left(1 - \Lambda_S\left[\hat{q}_S(x) - 1\right]\right) - \tilde{q}_S(x)\left(1 - \Lambda_S\left[\tilde{q}_S(x) - 1\right]\right)\right)\ dx \\
&\leq (2p+a)\int_{\underline{p}}^{\bar{p}} \left(\hat{q}_S(x)\left(1 - \Lambda_S\left[\hat{q}_S(x) - 1\right]\right) - \tilde{q}_S(x)\left(1 - \Lambda_S\left[\tilde{q}_S(x) - 1\right]\right)\right)\ dx
\end{aligned}$$

because the difference in the brackets is positive until p and then negative. Rewrite this difference to obtain

$$\begin{aligned}
&\int_{\underline{p}}^{\bar{p}} \left(\hat{q}_S(x)\left(1 - \Lambda_S\left[\hat{q}_S(x) - 1\right]\right) - \tilde{q}_S(x)\left(1 - \Lambda_S\left[\tilde{q}_S(x) - 1\right]\right)\right)\ dx \\
&= (1+\Lambda_S)\int_{\underline{p}}^{\bar{p}} \left(\hat{q}_S(x) - \tilde{q}_S(x)\right)\ dx + \Lambda_S \int_{\underline{p}}^{\bar{p}} \left(\tilde{q}_S(x) - \hat{q}_S(x)\right)\left(\hat{q}_S(x) + \tilde{q}_S(x)\right)\ dx.
\end{aligned}$$

The first integral is equal to zero by construction. In the second integral, note that the first bracket is negative until p and then positive and the second bracket is a decreasing function. Thus,

$$\begin{aligned}
&\Lambda_S \int_{\underline{p}}^{\bar{p}} \left(\tilde{q}_S(x) - \hat{q}_S(x)\right)\left(\hat{q}_S(x) + \tilde{q}_S(x)\right)\, dx \\
&\leq \left(\hat{q}_S(p) + \tilde{q}_S(p)\right)\Lambda_S \int_{\underline{p}}^{\bar{p}} \left(\tilde{q}_S(x) - \hat{q}_S(x)\right)\, dx \\
&= 0.
\end{aligned}$$

Overall, we have showed that

$$\int_0^1 (2x+a)\left(\hat{q}_S(x)\left(1 - \Lambda_S\left[\hat{q}_S(x) - 1\right]\right) - \tilde{q}_S(x)\left(1 - \Lambda_S\left[\tilde{q}_S(x) - 1\right]\right)\right)\, dx \leq 0$$

proving that $\tilde{q}_S$ was not a minimizer and that q_S^* indeed is the solution to the problem.

Step 4. Having eliminated the seller's interim trade probability from the problem using the allocation and aggregate constraints, the maximization problem reads

$$\begin{aligned}
&\max_{q_B} \int_0^1 (2x - 1 + a) q_B(x)\left(1 + \Lambda_B\left[q_B(x) - 1\right]\right)\, dx \\
&- \int_0^1 (2x+a)(1 - q_B^{-1}(x))\left(1 + \Lambda_S q_B^{-1}(x)\right)\, dx
\end{aligned}$$

subject to q_B being non-decreasing. We now use the substitution $x = q_B(t)$ to eliminate q_B^{-1} from the problem and obtain

$$\int_0^1 (2x - 1 + a) q_B(x)\left(1 + \Lambda_B\left[q_B(x) - 1\right]\right)\, dx - \int_{q_B^{-1}(0)}^{q_B^{-1}(1)} (2y_B(x) + a)(1-x)\left(1 + \Lambda_S x\right) q_B'(x)\, dx$$

Note that q_B is differentiable almost everywhere and therefore the substitution is well-defined. We will guess and verify that $y_B^{-1}(0) = 0$ and $y_B^{-1}(1) = 1$. The objective then becomes

$$\int_0^1 (2x - 1 + a) q_B(x)\left(1 + \Lambda_B\left[q_B(x) - 1\right]\right) - (2q_B(x) + a)(1-x)\left(1 + \Lambda_S x\right) y_B'(x)\, dx.$$

We perform one final substitution to ensure the positivity of q_B and let $q_B(t) = u^2(t)$. We then obtain

$$\begin{aligned}
&\int_0^1 (2x - 1 + a) u^2(x)\left(1 + \Lambda_B\left[u^2(x) - 1\right]\right) - (2u^2(x) + a)(1-x)\left(1 + \Lambda_S x\right) 2u(x)u'(x)\, dx \\
&= \int_0^1 J(x, u, u')\, dx.
\end{aligned}$$

We know from methods of calculus of variations that a necessary condition for a solution to the problem is characterized by

$$\frac{d}{dx} J_{u'}(x, u, u') = J_u(x, u, u').$$

We obtain the candidates for a maximum given by

$$u(x) = 0 \text{ and } u(x) = \pm\frac{\sqrt{(2x-1)(1-\Lambda_B) + 2a^2\Lambda_S - a((2x-1)\Lambda_S + 2 - \Lambda_B)}}{\sqrt{2(1-(2x-1-a)\Lambda_B + (2x-1-2a)\Lambda_S)}}$$

where the second candidate is only well-defined for all

$$x \geq \bar{x} = \frac{2a^2\Lambda_S + a\Lambda_B + a\Lambda_S - 2a + \Lambda_B - 1}{2(a\Lambda_S + \Lambda_B - 1)}.$$

Note that $x \leq a+1$ when $\Lambda_S \leq (1-\Lambda_B(a+1))/a$ and $\Lambda_B \leq 1/(1+a)$. Reversing the substitutions we obtain that the optimal interim trade probability for the buyer is

$$y_B^*(\theta_B) = \frac{2\theta_B(1-2\Lambda_B - 2\Lambda_S a) + 2a^2\Lambda_S + a(\Lambda_B + \Lambda_S - 2) + \Lambda_B - 1}{2(1-\Lambda_B(2\theta_B - 1 - a) + \Lambda_S(2\theta_B - 1 - 2a))}$$

if $\Lambda_S \leq (1-\Lambda_B(a+1))/a$ and $\Lambda_B \leq 1/(1+a)$ and $y_B^*(\theta_B) = 0$ otherwise. This interim trade probability (and the corresponding for the seller) can be obtained by the ex-post trade rule given by

$$y^{RM}(\theta_S, \theta_B) = \begin{cases} 1 & \text{if } \theta_S \leq \delta^{RM}(\theta_B), \\ 0 & \text{otherwise.} \end{cases}$$

If $\Lambda_S \leq (1-\Lambda_B(a+1))/a$ and $\Lambda_B \leq 1/(1+a)$, there exists $\bar{\theta}_B \in [a, a+1]$ such that $\delta^{RM}(\theta_B) = a$ for $\theta_B < \bar{\theta}_B$, and

$$\delta^{RM}(\theta_B) = \frac{(2\theta_B - 1 - a)(1-\Lambda_B(2a+1) + a\Lambda_S) + a - \Lambda_S a^2}{2(1-\Lambda_B(2\theta_B - a - 1) + \Lambda_S(2\theta_B - 1 - 2a))},$$

for $\theta_B \geq \bar{\theta}_B$. If $\Lambda_S > (1-\Lambda_B(a+1))/a$ or $\Lambda_B > 1/(1+a)$ we have $\delta^{RM}(\theta_B) = a$ for all $\theta_B \in [a, a+1]$.

Step 5. One can easily verify that the IR constraints are satisfied.

A.1.3 Maximizing the Gains from Trade

The derivations of the mechanisms maximizing the total and the material gains from trade proceed analogously. We here present the derivations for the case of maximizing the total gains from trade.

Step 1. We consider the problem of maximizing the total gains from trade. The analysis for the problem of maximizing the material gains from trade is analogous. We first rewrite the problem as a function of the trade rule only. We can rewrite the objective function to (imposing

$\Lambda_B = \Lambda_S = \Lambda$)

$$\begin{aligned}
&\int_a^b U_B(\theta_B, s_S^t|\theta_B)\, d\theta_B + \int_a^b U_S(\theta_S, s_B^t|\theta_S)\, d\theta_S \\
=&\int_a^b \Big(\theta_B y_B(\theta_B)(1+\Lambda(y_B(\theta_B)-1)) - \bar{t}_B(\theta_B) + \eta_B^2 w_B(\theta_B)\Big)\, d\theta_B \\
&- \int_a^b \Big(\theta_S y_S(\theta_S)(1-\Lambda(y_S(\theta_S)-1)) - \bar{t}_S(\theta_S) - \eta_S^2 w_S(\theta_S)\Big)\, d\theta_S.
\end{aligned}$$

Note that by the budget constraint (AB) we have

$$\int_a^b t_B(\theta_B)\, d\theta_B = \int_a^b t_S(\theta_S)\, d\theta_S.$$

Further, $w_B(\theta_B)$ and $w_S(\theta_S)$ enter the objective positively and both are negative. Hence, we optimally set both terms to zero by choosing interim deterministic transfers. This yields

$$\begin{aligned}
&\int_a^b U_B(\theta_B, s_S^t|\theta_B)\, d\theta_B + \int_a^b U_S(\theta_S, s_B^t|\theta_S)\, d\theta_S \\
=&\int_a^b \theta_B y_B(\theta_B)(1+\Lambda(y_B(\theta_B)-1))\, d\theta_B - \int_a^b \theta_S y_S(\theta_S)(1-\Lambda(y_S(\theta_S)-1))\, d\theta_S.
\end{aligned}$$

Mirroring the arguments in the proof of the revenue maximizing mechanism, the budget constraint AB and the CPEIC can be jointly written as

$$\begin{aligned}
&\int_a^b (2\theta_B - 1 - a) y_B(\theta_B)\left(1+\Lambda\left[y_B(\theta_B)-1\right]\right)\, d\theta_B \\
=&\int_a^b (2\theta_S - a) y_S(\theta_S)\left(1-\Lambda\left[y_S(\theta_S)-1\right]\right)\, d\theta_S,
\end{aligned}$$

as well as the monotonicity constraints. Thus, the maximization problem is a function of the trade rule only.

Step 2. We can set up the Lagrangian as

$$\begin{aligned}
\mathcal{L}(y^f, \gamma) = &\int_a^b (\theta_B + \gamma(2\theta_B - 1 - a)) y_B(\theta_B)\left(1+\Lambda\left[y_B(\theta_B)-1\right]\right)\, d\theta_B \\
&- \int_a^b (\theta_S + \gamma(2\theta_S - a)) y_S(\theta_S)\left(1-\Lambda\left[y_S(\theta_S)-1\right]\right)\, d\theta_S.
\end{aligned}$$

Note that we must have $\gamma \geq 0$, because relaxing the budget constraint (i.e., allowing the designer to run a deficit) can only increase the objective. Hence, $(\theta_B + \gamma(2\theta_B - 1 - a))$ and $(\theta_S + \gamma(2\theta_S - a))$ are strictly increasing in θ_B and θ_S, respectively. Therefore, the arguments in the proof of the revenue maximizing mechanism carry through and we can again maximize over the interim trade probabilities directly and eliminate y_S from the problem.

Step 3. Mirroring the steps in the proof of the revenue maximizing mechanism we obtain an expression for the interim trade probability of the buyer. Using the budget constraint and the assumption that $\Lambda = \Lambda_B = \Lambda_S$ we can eliminate the Lagrange multiplier from this expression. Next, reversing the change in variables we obtain the buyer's interim trade probability from

which we can recover the ex-post allocation rule which gives rise to the interim trade probabilities and is given by

$$y^{TG}(\theta_S, \theta_B) = \begin{cases} 1 & \text{if } \theta_S \leq \delta^{TG}(\theta_B), \\ 0 & otherwise. \end{cases}$$

If $\Lambda < 1/(1+a)$, there exists $\bar{\theta}_B \in [a, a+1]$ such that $\delta^{TG}(\theta_B) = a$ for $\theta_B < \bar{\theta}_B$, and

$$\begin{aligned} \delta^{TG}(\theta_B) = &\frac{(2a\Lambda + \Lambda - 1)((2a^2 + 2a + 1)\Lambda^2 - M - (2a+1)\Lambda)}{2(a\Lambda - 1)(M + a\Lambda^2 - (a+1)\Lambda + 1)} \\ &+ \theta_B \frac{(a\Lambda + \Lambda - 1)(M - a(\Lambda+1)\Lambda - \Lambda^2 + 1)}{(a\Lambda - 1)(M + a\Lambda^2 - (a+1)\Lambda + 1)}, \end{aligned}$$

for $\theta_B \geq \bar{\theta}_B$, where

$$M = \sqrt{(3a^2 + 3a + 1)\Lambda^4 - (2a+1)\Lambda^3 + a(a+1)\Lambda^2 - (2a+1)\Lambda + 1}.$$

If $\Lambda \geq 1/(1+a)$ we have $\delta^{WM}(\theta_B) = a$ for all $\theta_B \in [a, a+1]$. The optimality of no trade for large enough stakes follows directly from the revenue maximizing mechanism. We know from there that for $\Lambda \leq 1/(1+a)$ any mechanism which induces trade yields a negative expected revenue. Hence, any mechanism which induces trade violates the budget balance constraint. Consequently, for $\Lambda \leq 1/(1+a)$ no trade is the only feasible welfare maximizing mechanism.
Step 4. One can easily verify that the IR constraints are satisfied.

B Appendix: Chapter 2

B.1 Proofs

B.1.1 Proof of Lemma 2.1

(i) Suppose that $B(\succeq, f) \subseteq B(\succeq, f')$ holds for each $\succeq \in P(\Lambda)$. To show that $[f]_\Lambda N(\Lambda)[f']_\Lambda$, we proceed by contradiction and assume that there exist $\succeq \in P(\Lambda)$ and $S \subseteq X$ for which $c(d(\succeq, f), S) = x$ and $c(d(\succeq, f'), S) = y$ with $x \neq y$ and $y \succeq x$. The definition of c implies $(x, y) \in d(\succeq, f)$ and $(x, y) \notin d(\succeq, f')$. Together with $(x, y) \notin \succeq$ this implies $(x, y) \in B(\succeq, f)$ but $(x, y) \notin B(\succeq, f')$, a contradiction. For the converse, suppose that there exist $\succeq \in P(\Lambda)$ and $x, y \in X$ with $(x, y) \in B(\succeq, f)$ but $(x, y) \notin B(\succeq, f')$, which requires $x \neq y$. This implies $(x, y) \in d(\succeq, f)$ and $(x, y) \notin \succeq$, hence $(x, y) \notin d(\succeq, f')$. Then $c(d(\succeq, f'), \{x, y\}) = y \succeq x = c(d(\succeq, f), \{x, y\})$, which implies that $[f]_\Lambda N(\Lambda)[f']_\Lambda$ does not hold, by Definition 2.1.

(ii) Reflexivity and transitivity of $N(\Lambda)$ follow from the set inclusion characterization in statement (i). To show antisymmetry, consider any $f, f' \in F$ with $[f]_\Lambda N(\Lambda)[f']_\Lambda$ and $[f']_\Lambda N(\Lambda)[f]_\Lambda$. By (i) this is equivalent to $B(\succeq, f) = B(\succeq, f')$ and thus $d(\succeq, f) = d(\succeq, f')$ for each $\succeq \in P(\Lambda)$, hence $[f]_\Lambda = [f']_\Lambda$.

B.1.2 Proof of Proposition 2.1

Suppose $\succeq$ is identifiable, which implies that $\bar{\Lambda}(\succeq)$ is not identical to $\bar{\Lambda}(\succeq')$ for any other $\succeq'$. Then $P(\bar{\Lambda}(\succeq)) = \{\succeq\}$. Consider any f with $d(\succeq, f) = \succeq$, which exists by assumption. For any $f' \in F$, we then have $B(\succeq, f) = \varnothing \subseteq B(\succeq, f')$ and hence $[f]_{\bar{\Lambda}(\succeq)} N(\bar{\Lambda}(\succeq))[f']_{\bar{\Lambda}(\succeq)}$ by Lemma 2.1, which implies $f \in G(\bar{\Lambda}(\succeq))$. For the converse, suppose that $\succeq$ is not identifiable, i.e., there exists $\succeq' \neq \succeq$ with $\bar{\Lambda}(\succeq') = \bar{\Lambda}(\succeq)$. Then $\{\succeq, \succeq'\} \subseteq P(\bar{\Lambda}(\succeq))$. Consider any f_1 with $d(\succeq, f_1) = \succeq$ and any f_2 with $d(\succeq', f_2) = \succeq'$, so that $B(\succeq, f_1) = \varnothing$ and $B(\succeq', f_2) = \varnothing$. Assume by contradiction that there exists $f \in G(\bar{\Lambda}(\succeq))$. Then $[f]_{\bar{\Lambda}(\succeq)} N(\bar{\Lambda}(\succeq))[f_1]_{\bar{\Lambda}(\succeq)}$ must hold, which implies $B(\succeq, f) = \varnothing$ by Lemma 2.1, and hence $d(\succeq, f) = \succeq$. The analogous argument for f_2 implies $d(\succeq', f) = \succeq'$, which contradicts that $\bar{\Lambda}(\succeq') = \bar{\Lambda}(\succeq)$, i.e., that $\succeq$ is not identifiable.

B.1.3 Proof of Proposition 2.2

Any behavioral model d is characterized by the collection of maximal data sets $(\bar{\Lambda}(\succeq))_{\succeq \in P}$ that it assigns to the welfare preferences. Suppose there are $m_P \geq 2$ preferences and $m_F \geq 2$ frames. Then there are $(m_P)^{m_F}$ different maximal data sets. For a given welfare preference $\succeq$, however, only

$$N(m_P, m_F) = (m_P)^{m_F} - (m_P - 1)^{m_F}$$

of them are admissible, as the others contradict the existence of a non-distorting frame for $\succeq$. The number of possible models is thus given by $N(m_P, m_F)^{m_P}$. To obtain a model with identifiable preferences, we need to assign a different maximal data set to each welfare preference. Suppose we assign one of the $N(m_P, m_F)$ admissible data sets to the first welfare preference. Then there remain at least $N(m_P, m_F)-1$ admissible data sets for the second welfare preference (the exact number is still $N(m_P, m_F)$ if the data set assigned to the first preference was not admissible for the second preference), and so on. Observe that $N(m_P, m_F) \geq m_P$, so we can proceed iteratively and obtain the falling factorial

$$N(m_P, m_F)^{\underline{m_P}} = N(m_P, m_F) \times (N(m_P, m_F) - 1) \times \ldots \times (N(m_P, m_F) - m_P + 1)$$

as a lower bound on the number of models with identifiable preferences. Consequently,

$$S(m_P, m_F) = \frac{N(m_P, m_F)^{\underline{m_P}}}{N(m_P, m_F)^{m_P}}$$

is a lower bound on the share of models with identifiable preferences. We can rewrite

$$\begin{aligned} S(m_P, m_F) &= \frac{N(m_P, m_F)}{N(m_P, m_F)} \times \frac{N(m_P, m_F) - 1}{N(m_P, m_F)} \times \cdots \times \frac{N(m_P, m_F) - m_P + 1}{N(m_P, m_F)} \\ &= \prod_{k=1}^{m_P - 1} \left(1 - \frac{k}{N(m_P, m_F)}\right) \\ &= \exp\left(\sum_{k=1}^{m_P - 1} \log\left(1 - \frac{k}{N(m_P, m_F)}\right)\right), \end{aligned}$$

where $1 > k/N(m_P, m_F) > 0$ holds for all $k = 1, \ldots, m_P - 1$. Recall that for $x > -1$ we have $\log(1 + x) \geq x/(1 + x)$, which implies

$$\sum_{k=1}^{m_P - 1} \log\left(1 - \frac{k}{N(m_P, m_F)}\right) \geq \sum_{k=1}^{m_P - 1} -\frac{k}{N(m_P, m_F) - k}.$$

Furthermore,

$$\begin{aligned} \sum_{k=1}^{m_P - 1} -\frac{k}{N(m_P, m_F) - k} &\geq \sum_{k=1}^{m_P - 1} -\frac{k}{N(m_P, m_F) - m_P + 1} \\ &= -\frac{1}{N(m_P, m_F) - m_P + 1} \sum_{k=1}^{m_P - 1} k \\ &= -\frac{(m_P)^2 - m_P}{2(N(m_P, m_F) - m_P + 1)}. \end{aligned}$$

Altogether, we therefore have

$$S(m_P, m_F) \geq \exp\left(-\frac{(m_P)^2 - m_P}{2(N(m_P, m_F) - m_P + 1)}\right) = \tilde{S}(m_P, m_F),$$

so $\tilde{S}(m_P, m_F)$ is also a lower bound on the share of models with identifiable preferences.

We are interested in asymptotic behavior as the number of alternatives m_X and hence the number of preferences m_P grows. Holding m_F fixed and treating m_P as a real variable, it follows with l'Hôpital's rule that

$$\lim_{m_P \to \infty} -\frac{(m_P)^2 - m_P}{2(N(m_P, m_F) - m_P + 1)} = 0$$

whenever $m_F \geq 4$. We thus obtain $\lim_{m_X \to \infty} \tilde{S}(m_P(m_X), m_F) = 1$ whenever $m_F \geq 4$. Now consider the case that the number of frames $m_F(m_X)$ also depends on the number of alternatives. Observe that $\tilde{S}(m_P, m_F)$ is strictly increasing in m_F whenever $m_P \geq 2$. At the same time, $\tilde{S}(m_P, m_F) \leq 1$ always holds since $\tilde{S}(m_P, m_F)$ is a lower bound on a proportion. Hence we obtain

$$\lim_{m_X \to \infty} \tilde{S}(m_P(m_X), m_F(m_X)) = 1$$

whenever there exists $\underline{m}$ such that $m_F(m_X) \geq 4$ for all $m_X \geq \underline{m}$. Then the share of models with identifiable preferences converges to 1 as the number of alternatives grows to infinity.

B.1.4 Proof of Proposition 2.3

Consider any d with the frame-cancellation property and any data set Λ. Fix any frame $f_1 \in F$, and let $f_2 \in F$ be an arbitrary frame with $f_2 \notin [f_1]_\Lambda$. Then, by definition of $[f_1]_\Lambda$, there exists $\succeq \in P(\Lambda)$ such that $d(\succeq, f_1) = \succeq_1 \neq \succeq_2 = d(\succeq, f_2)$. By the frame-cancellation property, we have $d(\succeq_1, f) = d(d(\succeq, f_1), f) = d(\succeq, f)$ for all $f \in F$, which implies that $\succeq_1 \in P(\Lambda)$. We also obtain $d(\succeq_1, f_1) = d(\succeq, f_1) = \succeq_1$, which implies $B(\succeq_1, f_1) = \varnothing$. From $\succeq_1 \neq \succeq_2$ and the frame-cancellation property, it follows that

$$B(\succeq_1, f_2) = d(\succeq_1, f_2) \setminus \succeq_1 = d(d(\succeq, f_1), f_2) \setminus \succeq_1 = d(\succeq, f_2) \setminus \succeq_1 = \succeq_2 \setminus \succeq_1 \neq \varnothing.$$

Hence $B(\succeq_1, f_1) \subset B(\succeq_1, f_2)$, and Lemma 2.1 implies that $[f_2]_\Lambda N(\Lambda)[f_1]_\Lambda$ does not hold. Since f_2 was arbitrary we conclude that $f_1 \in M(\Lambda)$, and, since f_1 was arbitrary, that $M(\Lambda) = F$.

B.1.5 Proof of Proposition 2.4

We assume $k \leq m_X/2$ throughout the proof, as cases where $k > m_X/2$ can be dealt with equivalently by reversing the role of the first page f and the second page $X \setminus f$ of the search engine.

Case 1: k even. We first construct an elicitation procedure e and then show that it is optimal. Let $e(\varnothing) = f_1$ be an arbitrary subset $f_1 \subseteq X$ with $|f_1| = k$. Now fix any welfare preference $\succeq$. The procedure then generates a data set $\Lambda_1 = \{(\succeq_1, f_1)\} \in L_1$, where $\succeq_1$ agrees with $\succeq$ within the sets f_1 and $X \setminus f_1$. Let a_i denote the alternative ranked at position i within the set f_1 by $\succeq_1$, for each $i = 1, \ldots, k$. Let b_i denote the alternative ranked at position i within the set $X \setminus f_1$ by $\succeq_1$, for each $i = 1, \ldots, k, \ldots, m_X - k$. Then construct the frame $e(\Lambda_1) = f_2$ as $f_2 = \{a_1, \ldots, a_{k/2}, b_{k/2+1}, \ldots, b_k\}$. The procedure then generates a data set $\Lambda_2 = \{(\succeq_1, f_1), (\succeq_2, f_2)\} \in L_2$, where $\succeq_2$ agrees with $\succeq$ within the sets f_2 and $X \setminus f_2$. This construction is applied to all the data sets Λ_1 that are generated by the elicitation procedure

for some welfare preference. The elicitation procedure can be continued arbitrarily for all other data sets.

Let $\succeq$ be an arbitrary true welfare preference. We claim that the set $T_k(\succeq)$ of top k alternatives according to $\succeq$ can be deduced from the generated Λ_2, so that the optimal nudge is identified and $n(e,\succeq) \leq 2$ follows. Observe first that none of the alternatives $b_{k+1},\dots,b_{m_X-k}$ (if they exist) can belong to $T_k(\succeq)$, because Λ_1 has already revealed that each $b_1,\dots,b_k$ is preferred by $\succeq$. Now suppose that $b_k \succeq_2 a_1$ holds. We then know that $b_k \succeq a_1$ and thus $T_k(\succeq) = \{b_1,\dots,b_k\}$. Otherwise, if $a_1 \succeq_2 b_k$ holds, we know that $a_1 \succeq b_k$ and thus $b_k \notin T_k(\succeq)$ but $a_1 \in T_k(\succeq)$. In this case we can repeat the argument for a_2 and b_{k-1}: if $b_{k-1} \succeq_2 a_2$ we know that $b_{k-1} \succeq a_2$ and thus $T_k(\succeq) = \{b_1,\dots,b_{k-1},a_1\}$; otherwise, if $a_2 \succeq_2 b_{k-1}$ holds, we know that $a_2 \succeq b_{k-1}$ and thus $b_{k-1} \notin T_k(\succeq)$ but $a_2 \in T_k(\succeq)$. Iteration either reveals $T_k(\succeq)$ or arrives at $a_{k/2} \succeq_2 b_{k/2+1}$, which implies $a_{k/2} \succeq b_{k/2+1}$. In this case, we know that $T_k(\succeq)$ consists of $a_1,\dots,a_{k/2}$ and those $k/2$ alternatives that $\succeq_2$ and hence $\succeq$ ranks top within $X\backslash f_2$.

Since $\succeq$ was arbitrary, we know that $\max_{\succeq\in P} n(e,\succeq) \leq 2$. Obviously, no single observation ever suffices to deduce $T_k(\succeq)$, neither in the constructed procedure nor in any other one, hence we can conclude that $n = 2$.

Case 2: k odd and $k < m_X/2$. The construction is the same as for case 1, except that $f_2 = \{a_1,\dots,a_{(k-1)/2},b_{(k+1)/2+1},\dots,b_k,b_{k+1}\}$, where b_{k+1} exists because $k < m_X/2$. The arguments about deducing $T_k(\succeq)$ are also the same, starting with a comparison of a_1 and b_k, except that the iteration might arrive at $a_{(k-1)/2} \succeq_2 b_{(k+1)/2+1}$, in which case $T_k(\succeq)$ consists of $a_1,\dots,a_{(k-1)/2}$ and those $(k+1)/2$ alternatives that $\succeq_2$ ranks top within $X\backslash f_2$.

Case 3: k odd and $k = m_X/2$. The construction is the same as for case 1, except that $f_2 = \{a_1,\dots,a_{(k+1)/2},b_{(k+1)/2+1},\dots,b_k\}$. The arguments about deducing $T_k(\succeq)$ are also the same, starting with a comparison of a_1 and b_k, except that the iteration might arrive at $a_{(k-1)/2} \succeq_2 b_{(k+1)/2+1}$. In this case, we can conclude that $T_k(\succeq)$ consists of $a_1,\dots,a_{(k-1)/2}$, plus either $a_{(k+1)/2}$ or $b_{(k+1)/2}$ but never both, and those $(k-1)/2$ alternatives that $\succeq_2$ ranks top among the remaining ones in $X\backslash f_2$. Hence there exist welfare preferences $\succeq$ for which e does not identify $T_k(\succeq)$ after two steps. Since the missing preference between $a_{(k+1)/2}$ and $b_{(k+1)/2}$ can be learned by having $e(\Lambda_2) = f_3$ satisfy $\{a_{(k+1)/2},b_{(k+1)/2}\} \subseteq f_3$, we know that $n \leq 3$.

It remains to be shown that $n > 2$. Fix an arbitrary elicitation procedure e and denote $e(\varnothing) = f_1 = \{a_1,\dots,a_k\}$ and $X\backslash f_1 = \{b_1,\dots,b_k\}$, where the numbering of the alternatives is arbitrary but fixed (remember that $k = m_X/2$). Let $\succeq_1$ be the preference given (in ranking notation) by $a_1 \dots a_k\, b_1 \dots b_k$, and consider the data set $\Lambda_1 = \{(\succeq_1, f_1)\}$ and the subsequent frame $e(\Lambda_1) = f_2$. Since k is odd, it follows that at least one of the pairs $\{a_1,b_k\},\{a_2,b_{k-1}\},\dots,\{a_k,b_1\}$ must be separated on different pages by f_2, i.e., there exists $l = 1,\dots,k$ such that $a_l \in f_2$ and $b_{k-l+1} \in X\backslash f_2$ or vice versa. Depending on the value of l, we now construct two welfare preferences $\succeq'$ and $\succeq''$. If $l = 1$, let

$$\succeq'\colon\ b_1 \dots b_{k-1}\, b_k\, a_1\, a_2 \dots a_k,$$
$$\succeq''\colon\ b_1 \dots b_{k-1}\, a_1\, b_k\, a_2 \dots a_k.$$

If $l = 2, \ldots, k-1$, let

$$\begin{aligned} \succeq'\colon\ & a_1 \ldots a_{l-1}\, b_1 \ldots b_{k-l}\, b_{k-l+1}\, a_l\, a_{l+1} \ldots a_k\, b_{k-l+2} \ldots b_k, \\ \succeq''\colon\ & a_1 \ldots a_{l-1}\, b_1 \ldots b_{k-l}\, a_l\, b_{k-l+1}\, a_{l+1} \ldots a_k\, b_{k-l+2} \ldots b_k. \end{aligned}$$

If $l = k$, let

$$\begin{aligned} \succeq'\colon\ & a_1 \ldots a_{k-1}\, b_1\, a_k\, b_2 \ldots b_k, \\ \succeq''\colon\ & a_1 \ldots a_{k-1}\, a_k\, b_1\, b_2 \ldots b_k. \end{aligned}$$

For the two constructed welfare preferences $\succeq'$ and $\succeq''$, the elicitation procedure first generates the above described data set Λ_1. Subsequently, it generates the same data set $\Lambda_2 = \{(\succeq_1, f_1), (\succeq_2, f_2)\}$, because $\succeq'$ and $\succeq''$ differ only with respect to a_l and b_{k-l+1}, which is not revealed by frame f_2. Since $T_k(\succeq') \neq T_k(\succeq'')$, it follows that $n(e, \succeq') > 2$, which implies $\max_{\succeq \in P} n(e, \succeq) > 2$. Since e was arbitrary, it follows that $n > 2$.

B.1.6 Proof of Proposition 2.5

The result follows immediately if $m_X = 2$. Hence we fix a set X with $m_X \geq 3$ throughout the proof. We denote $m = m_X!$ for convenience.

Consider an arbitrary behavioral model, given by F and d, with $m_F \geq m$ and identifiable preferences. Define

$$\hat{n}(e, \succeq) = \min\{s \mid P(\Lambda_s(e, \succeq)) = \{\succeq\}\}$$

as the first step at which procedure e identifies $\succeq$, and let

$$\hat{n} = \min_{e \in E} \max_{\succeq \in P} \hat{n}(e, \succeq).$$

It follows immediately that $n \leq \hat{n}$, because $P(\Lambda_s(e, \succeq)) = \{\succeq\}$ implies $G(\Lambda_s(e, \succeq)) \neq \varnothing$. We will establish the inequality $\hat{n} < m$.

Consider any e and suppose $\hat{n}(e, \succeq) \geq m$ for some $\succeq \in P$. Since $|P| = m$, there must exist $k \in \{0, 1, \ldots, m-2\}$ such that

$$P(\Lambda_k(e, \succeq)) = P(\Lambda_{k+1}(e, \succeq)).$$

Denoting $e(\Lambda_k(e, \succeq)) = \tilde{f}$ and $d(\succeq, \tilde{f}) = \tilde{\succeq}$, we thus have $\Lambda_{k+1}(e, \succeq) = \Lambda_k(e, \succeq) \cup \{(\tilde{\succeq}, \tilde{f})\}$ and $d(\succeq', \tilde{f}) = \tilde{\succeq}$ for all $\succeq' \in P(\Lambda_k(e, \succeq))$. We now define elicitation procedure e' by letting $e'(\Lambda) = e(\Lambda)$, except for data sets $\Lambda \in L$ that satisfy both $\Lambda_k(e, \succeq) \subseteq \Lambda$ and $f \neq \tilde{f}$ for all $(\succeq, f) \in \Lambda$, which includes $\Lambda = \Lambda_k(e, \succeq)$. For those data sets, we define

$$e'(\Lambda) = \begin{cases} e(\Lambda \cup \{(\tilde{\succeq}, \tilde{f})\}) & \text{if} \quad |\Lambda| \leq m_F - 2, \\ \tilde{f} & \text{if} \quad |\Lambda| = m_F - 1. \end{cases}$$

Note that e' is a well-defined elicitation procedure. First, $\Lambda \cup \{(\tilde{\succeq}, \tilde{f})\} \in L$ holds whenever the first case applies, because $\varnothing \neq P(\Lambda) \subseteq P(\Lambda_k(e, \succeq))$ and Λ does not yet contain an observation of $\tilde{f}$. Second, the first case then applies repeatedly because $e(\Lambda \cup \{(\tilde{\succeq}, \tilde{f})\}) \neq \tilde{f}$, so that e' only dictates yet unobserved frames.

Consider any $\succeq' \notin P(\Lambda_k(e,\succeq))$, so that $(\succeq_1, f) \in \Lambda_k(e,\succeq')$ and $(\succeq_2, f) \in \Lambda_k(e,\succeq)$ with $\succeq_1 \neq \succeq_2$ for some f. From $\Lambda_k(e,\succeq') \subseteq \Lambda_s(e,\succeq')$ and thus $\Lambda_k(e,\succeq) \nsubseteq \Lambda_s(e,\succeq')$ for all $s \geq k$, it follows that preference $\succeq'$ is unaffected by the modification of the procedure, i.e., $\Lambda_s(e',\succeq') = \Lambda_s(e,\succeq')$ for all $s \in \{0,1,\ldots,m_F\}$, so that $\hat{n}(e',\succeq') = \hat{n}(e,\succeq')$. Now consider any $\succeq' \in P(\Lambda_k(e,\succeq))$, including $\succeq' = \succeq$. Then $\Lambda_s(e,\succeq) = \Lambda_s(e,\succeq') = \Lambda_s(e',\succeq')$ holds for all $s \leq k$. For $k < s \leq m_F - 1$, the definition of e' implies that $\Lambda_s(e',\succeq')$ does not contain an observation of $\tilde{f}$, and that

$$\Lambda_s(e',\succeq') \cup \{(\tilde{\succeq}, \tilde{f})\} = \Lambda_{s+1}(e,\succeq').$$

Thus

$$P(\Lambda_s(e',\succeq')) = P(\Lambda_s(e',\succeq') \cup \{(\tilde{\succeq}, \tilde{f})\}) = P(\Lambda_{s+1}(e,\succeq')),$$

so that $\hat{n}(e',\succeq') = \hat{n}(e,\succeq') - 1$. Repeated application of this construction allows us to arrive at an elicitation procedure e^* for which $\hat{n}(e^*,\succeq) < m$ for all $\succeq \in P$, which implies that $\hat{n} < m$.

B.1.7 Proof of Proposition 2.6

We write $P = \{\succeq_1, \succeq_2, \ldots, \succeq_m\}$, where $m = m_X!$ and the numbering of the preferences is arbitrary but fixed. We number the frames such that $f_i = o(\succeq_i)$. Note that each frame f_i is non-distorting for a single preference only, the one with which it coincides. This implies $n(e,\succeq) = \hat{n}(e,\succeq)$ for all $e \in E$ and $\succeq \in P$, and thus $n = \hat{n}$, where $\hat{n}$ refers to the complexity of identifying the welfare preference as defined in the proof of Proposition 2.5. We will establish the equality $\hat{n} = m - 1$.

Consider an arbitrary e. Define i_1 such that $e(\varnothing) = f_{i_1}$, and i_t for $t = 2, 3, \ldots, m$ recursively such that $e(\Lambda_{t-1}) = f_{i_t}$ for the data set

$$\Lambda_{t-1} = \bigcup_{j=1}^{t-1} \{(f_{i_j}, f_{i_j})\}.$$

If $\succeq_{i_m}$ is the welfare preference, then the procedure e will generate the sequence of data sets $\Lambda_s(e,\succeq_{i_m}) = \Lambda_s$ for all $s \in \{0,1,\ldots,m-1\}$, with $\Lambda_0 = \varnothing$. It follows from the definition of d that $P(\Lambda_s) = \{\succeq_{i_{s+1}}, \succeq_{i_{s+2}}, \ldots, \succeq_{i_m}\}$ holds for each $s \in \{0,1,\ldots,m-1\}$. This implies $\hat{n}(e,\succeq_{i_m}) = m - 1$, and hence $\max_{\succeq \in P} \hat{n}(e,\succeq) \geq m - 1$. Since e was arbitrary, it follows that $\hat{n} \geq m - 1$. Together with the result $\hat{n} < m$ established in the proof of Proposition 2.5, this implies $\hat{n} = m - 1$.

B.1.8 Proof of Proposition 2.7

We first show that the partition $\bar{P}$ for the perfect-recall model, denoted $\bar{P}^{PR}$, is weakly finer than the one for the no-recall model, denoted $\bar{P}^{NR}$, and strictly so whenever $k < m_X - 1$. Fix any $\succeq \in P$ and consider the equivalence class $P^{PR}_{\succeq}$ in the perfect-recall model. For any $\succeq' \in P^{PR}_{\succeq}$ it holds that $T_k(\succeq') = T_k(\succeq)$, where $T_k(.)$ is the set of top k alternatives according to the respective preference. Hence $\succeq' \in P^{NR}_{\succeq}$, which implies that $P^{PR}_{\succeq} \subseteq P^{NR}_{\succeq}$ and hence that $\bar{P}^{PR}$ is weakly finer than $\bar{P}^{NR}$. If $k < m_X - 1$, we have $|X \backslash T_k(\succeq)| \geq 2$. Construct $\succeq''$ from $\succeq$ by swapping the preference between the bottom 2 alternatives. Then $T_k(\succeq'') = T_k(\succeq)$ and

hence $\succeq'' \in P^{NR}_{\succeq}$, but $\succeq'' \notin P^{PR}_{\succeq}$. Thus $\bar{P}^{PR}$ is strictly finer than $\bar{P}^{NR}$ in that case.

The fact that $\bar{P}^{PR}$ is weakly finer than $\bar{P}^{NR}$ immediately implies that any nudging domain for no-recall satisficing is also a nudging domain for perfect-recall satisficing. If $k < m_X - 1$, the domain $\tilde{P} = \{\succeq, \succeq''\}$ for two preferences $\succeq$ and $\succeq''$ as constructed above is a nudging domain for perfect-recall satisficing but not for no-recall satisficing.

B.1.9 Proof of Proposition 2.8

The proof is similar to the proofs of Propositions 2.5 and 2.6 and therefore omitted.

B.1.10 Proof of Proposition 2.9

As argued in the proof of Proposition 2.6, the strong priming model satisfies $n(e, \succeq) = \hat{n}(e, \succeq)$ for all $e \in E$ and $\succeq \in P$, where $\hat{n}(e, \succeq)$ denotes the first step at which procedure e identifies $\succeq$. Hence

$$\bar{n} = \min_{e \in E} \sum_{i=1}^{m} p_i \hat{n}(e, \succeq_i),$$

where we again write $m = m_X!$ for convenience. We also keep the numbering of frames such that $f_i = o(\succeq_i)$.

Consider an arbitrary e. Define i_t for $t = 1, 2, \ldots, m$ exactly as in the proof of Proposition 2.6, i.e., as the index of the frame prescribed by e at step t when the agent has been successfully manipulated by all previous frames. It then follows from the definition of d that $\hat{n}(e, \succeq_{i_t}) = t$ for each $t = 1, 2, \ldots, m-1$, and $\hat{n}(e, \succeq_{i_m}) = m - 1$. Hence

$$\sum_{i=1}^{m} p_i \hat{n}(e, \succeq_i) = \sum_{t=1}^{m} p_{i_t} \hat{n}(e, \succeq_{i_t}) = \sum_{t=1}^{m-1} p_{i_t} t + p_{i_m}(m-1),$$

which is a weighted average of the numbers $1, 2, \ldots, m-1, m-1$, where the weights are the prior probabilities. Since $p_1 \geq p_2 \geq \ldots \geq p_m$, this weighted average is minimized by a procedure $e \in E$ with $i_t = t$, which implies the result.

B.1.11 Proof of Proposition 2.10

Since each frame is non-distorting for exactly one preference in the strong priming model, we have $\bar{\varphi}_\Lambda = \max_{\succeq \in P(\Lambda)} \pi_\Lambda(\succeq)$. We now proceed in two steps. We first construct an elicitation procedure e, and then show that it is optimal.

Step 1. We only need to describe $e(\Lambda)$ for Λ with $|P(\Lambda)| \geq 2$, as otherwise $\bar{\varphi}_\Lambda = 1$ holds and the continuation of e is irrelevant for the generalized complexity. Given any such Λ, let j be the second-smallest index among the preferences in $P(\Lambda)$, so that $\pi_\Lambda(\succeq_j)$ is the second-highest value among the updated probabilities. Then we define $e(\Lambda) = f_j$ for this data set, where the numbering of frames is given by $f_i = o(\succeq_i)$ as before. Note that the frame f_j cannot have been observed in Λ already, since otherwise either $P(\Lambda) = \{\succeq_j\}$ or $\succeq_j \notin P(\Lambda)$ would hold. Hence the construction yields a well-defined elicitation procedure. For instance, we obtain $e(\varnothing) = f_2$, $e(\{(f_2, f_2)\}) = f_3$, and so on. If $\succeq_1$ is the welfare preference, it follows from the definition of d

that

$$P(\Lambda_k(e,\succeq_1)) = \begin{cases} \{\succeq_1, \succeq_{k+2}, \succeq_{k+3}, \ldots, \succeq_m\} & \text{if } k \leq m-2, \\ \{\succeq_1\} & \text{if } k \geq m-1, \end{cases}$$

and therefore

$$\bar{\varphi}_{\Lambda_k(e,\succeq_1)} = \begin{cases} p_1/(p_1 + \sum_{j=k+2}^m p_j) & \text{if } k \leq m-2, \\ 1 & \text{if } k \geq m-1, \end{cases} \tag{B.1}$$

where we once more write $m = m_X!$ for convenience. If $\succeq_i$ for $i = 2, 3, \ldots, m$ is the welfare preference, we have

$$P(\Lambda_k(e,\succeq_i)) = \begin{cases} \{\succeq_1, \succeq_{k+2}, \succeq_{k+3}, \ldots, \succeq_m\} & \text{if } k \leq i-2, \\ \{\succeq_i\} & \text{if } k \geq i-1, \end{cases}$$

and therefore

$$\bar{\varphi}_{\Lambda_k(e,\succeq_i)} = \begin{cases} p_1/(p_1 + \sum_{j=k+2}^m p_j) & \text{if } k \leq i-2, \\ 1 & \text{if } k \geq i-1. \end{cases} \tag{B.2}$$

Given any $k = 0, 1, \ldots, m$, the value of (B.1) is always weakly smaller than the value of (B.2). Hence $\max_{\succeq \in P} n(q, e, \succeq) = n(q, e, \succeq_1)$. The value of $n(q, e, \succeq_1)$ is given by the smallest integer $k \geq 0$ such that

$$\frac{p_1}{p_1 + \sum_{j=k+2}^m p_j} \geq q,$$

which can be rearranged to the condition in the proposition.

Step 2. Now consider an arbitrary elicitation procedure e. Define i_t for $t = 1, 2, \ldots, m$ exactly as in the proof of Proposition 2.6. For any $i = 1, 2, \ldots, m$ let $t(i)$ be such that $i = i_{t(i)}$, so that frame f_i is prescribed by e at step $t(i)$ when the agent has been successfully manipulated by all previous frames. We then obtain

$$P(\Lambda_k(e,\succeq_i)) = \begin{cases} \{\succeq_{i_j} | \, j = k+1, k+2, \ldots, m\} & \text{if } k \leq t(i)-1, \\ \{\succeq_i\} & \text{if } k \geq t(i), \end{cases}$$

and

$$\bar{\varphi}_{\Lambda_k(e,\succeq_i)} = \begin{cases} p_{i_{j^*(k)}}/(\sum_{j=k+1}^m p_{i_j}) & \text{if } k \leq t(i)-1, \\ 1 & \text{if } k \geq t(i), \end{cases} \tag{B.3}$$

where $j^*(k)$ is an index j in $\{k+1, k+2, \ldots, m\}$ for which p_{i_j} is maximal. Given any $k = 0, 1, \ldots, m$, the value of (B.3) is minimized when $t(i) = m$, i.e., for welfare preference $\succeq_i = \succeq_{i_m}$. Hence $\max_{\succeq \in P} n(q, e, \succeq) = n(q, e, \succeq_{i_m})$. We now claim that the value of (B.3) for $\succeq_{i_m}$ is weakly smaller than the value of (B.1), for all $k = 0, 1, \ldots, m$, from which it follows that the procedure constructed in step 1 is indeed optimal. We only need to establish the inequality

$$\frac{p_{i_{j^*(k)}}}{\sum_{j=k+1}^m p_{i_j}} \leq \frac{p_1}{p_1 + \sum_{j=k+2}^m p_j}$$

for all $k \leq m-2$. It can be rearranged to

$$p_{i_{j^*(k)}} \left(p_1 + \sum_{j \in \{k+2,\ldots,m\}} p_j \right) \leq p_1 \left(p_{i_{j^*(k)}} + \sum_{j \in \{k+1,k+2,\ldots,m\} \setminus \{j^*(k)\}} p_{i_j} \right),$$

which can further be rearranged to

$$\frac{\sum_{j \in \{k+2,\ldots,m\}} p_j}{\sum_{j \in \{k+1,k+2,\ldots,m\} \setminus \{j^*(k)\}} p_{i_j}} \leq \frac{p_1}{p_{i_{j^*(k)}}}.$$

This holds, because $p_1 \geq p_2 \geq \ldots \geq p_m$ implies that the LHS is weakly smaller than 1 while the RHS is weakly larger than 1.

B.1.12 Proof of Proposition 2.11

We only need to consider the case $\underline{q} < q \leq \overline{q}$, which presupposes the existence of a procedure e^* with which

$$\bar{\varphi}_{\varnothing} < q \leq \max_{s \in \{1,\ldots,m_F\}} \bar{\varphi}_{\Lambda_s(e^*,\succeq)}$$

for all $\succeq \in P$. We will show that, with the frame-cancellation property, for all $\succeq \in P$ it holds that $P(\Lambda_1(e^*,\succeq)) = P(\Lambda_s(e^*,\succeq))$ for all $s = 2,\ldots,m_F$, and therefore

$$\bar{\varphi}_{\Lambda_1(e^*,\succeq)} = \max_{s \in \{1,\ldots,m_F\}} \bar{\varphi}_{\Lambda_s(e^*,\succeq)}.$$

This then immediately implies $n(q) = 1$. We will in fact establish the stronger property that $P(\Lambda) = P(\Lambda')$ whenever $\varnothing \neq \Lambda \subseteq \Lambda'$.

We first show that, for any two $\succeq, \succeq' \in P$, the maximal data sets $\bar{\Lambda}(\succeq)$ and $\bar{\Lambda}(\succeq')$ are either disjoint or identical. Suppose $\bar{\Lambda}(\succeq)$ and $\bar{\Lambda}(\succeq')$ are not disjoint, so there exists $f' \in F$ such that $d(\succeq, f') = d(\succeq', f')$. Then the frame-cancellation property implies

$$d(\succeq, f) = d(d(\succeq, f'), f) = d(d(\succeq', f'), f) = d(\succeq', f)$$

for all $f \in F$, so that $\bar{\Lambda}(\succeq) = \bar{\Lambda}(\succeq')$.

Now fix any two data sets Λ and Λ' with $\varnothing \neq \Lambda \subseteq \Lambda'$. Since $P(\Lambda') \subseteq P(\Lambda)$ always holds, we only need to show that $P(\Lambda) \subseteq P(\Lambda')$. Fix any $\succeq \in P(\Lambda)$, so that $\Lambda \subseteq \bar{\Lambda}(\succeq)$. For any $\succeq' \in P(\Lambda')$ it holds that $\Lambda \subseteq \Lambda' \subseteq \bar{\Lambda}(\succeq')$. Since $\Lambda \neq \varnothing$, this implies that $\bar{\Lambda}(\succeq)$ and $\bar{\Lambda}(\succeq')$ are not disjoint, so that $\bar{\Lambda}(\succeq) = \bar{\Lambda}(\succeq')$. Hence $\Lambda' \subseteq \bar{\Lambda}(\succeq)$ and $\succeq \in P(\Lambda')$.

B.1.13 Proof of Proposition 2.12

We first show that, for all Λ and f, $P(\Lambda, f) = d(P(\Lambda), f)$ holds under the frame-cancellation property, where $d(P(\Lambda), f)$ denotes the image of $P(\Lambda)$ under $d(., f)$. The first inclusion $P(\Lambda, f) \subseteq d(P(\Lambda), f)$ follows immediately by definition of $d(P(\Lambda), f)$. Assume then that $\succeq \in d(P(\Lambda), f)$, so there exists $\succeq' \in P(\Lambda)$ such that $d(\succeq', f) = \succeq$. The frame-cancellation property then implies $d(\succeq, f') = d(d(\succeq', f), f') = d(\succeq', f')$ for any $f' \in F$, which reveals that $\succeq \in P(\Lambda)$. Furthermore, $d(\succeq, f) = d(\succeq', f) = \succeq$. Hence $\succeq \in P(\Lambda, f)$, so that the other inclusion $d(P(\Lambda), f) \subseteq P(\Lambda, f)$

also holds. We can therefore write

$$\varphi_\Lambda(f) = \sum_{\succeq \in d(P(\Lambda),f)} \pi_\Lambda(\succeq).$$

Consider $\Lambda = \varnothing$ first. For any $f \in F$, we claim that $|d(P(\varnothing), f)| = |d(P, f)| = |\bar{P}|$. Since f already partitions P into the $|d(P, f)|$ blocks between which it distinguishes, $|d(P, f)|$ is clearly a lower bound on $|\bar{P}|$. Now suppose $|d(P, f)| < |\bar{P}|$, which implies that there exist $\succeq_1 \neq \succeq_2$ such that $d(\succeq_1, f) = d(\succeq_2, f)$ but $d(\succeq_1, f') \neq d(\succeq_2, f')$ for some $f' \in F$. Then the frame-cancellation property implies $d(\succeq_1, f') = d(d(\succeq_1, f), f') = d(d(\succeq_2, f), f') = d(\succeq_2, f')$, a contradiction. Hence

$$\varphi_\varnothing(f) = \frac{|d(P, f)|}{m_X!} = \frac{|\bar{P}|}{m_X!} = \frac{1}{\bar{s}} \tag{B.4}$$

for all $f \in F$, i.e., initially each frame is equally likely to be optimal.

Consider any $\Lambda \neq \varnothing$ next. It has been shown in the proof of Proposition 2.11 that $P(\Lambda) = P(\Lambda')$ whenever $\varnothing \neq \Lambda \subseteq \Lambda'$, which implies that $P(\Lambda) = P(\bar{\Lambda}(\succeq)) = P_\succeq$ for any $\succeq \in P(\Lambda)$. For any $f \in F$, we then obtain that $|d(P(\Lambda), f)| = |d(P_\succeq, f)| = 1$ immediately from the definition of $P_\succeq$. Hence

$$\varphi_\Lambda(f) = \frac{1/m_X!}{s_\succeq/m_X!} = \frac{1}{s_\succeq} \tag{B.5}$$

for all $f \in F$, where $s_\succeq = |P_\succeq|$. Again, each frame is equally likely to be optimal.

Equation (B.4) also implies $\underline{q} = 1/\bar{s}$. Now consider any $\succeq' \in P$ with $s_{\succeq'} \geq \bar{s}$, which must clearly exist. For any procedure e and any $s = 1, \ldots, m_F$, we then have by equation (B.5)

$$\bar{\varphi}_{\Lambda_s(e,\succeq')} = \frac{1}{s_{\succeq'}} \leq \frac{1}{\bar{s}} = \underline{q},$$

which implies $\overline{q} = \underline{q}$.

B.1.14 Proof of Proposition 2.13

The proof is similar to the proof of Proposition 2.1 and therefore omitted.

B.1.15 Proof of Proposition 2.14

The proof is similar to the proof of Proposition 2.1 and therefore omitted.

B.1.16 Proof of Proposition 2.15

Case 1: $\delta \leq 1/(1+\overline{r})$. Fix any $(r, y) \in C$. If either $\delta < 1/(1+\overline{r})$ or $r < \overline{r}$, or both, we have $\delta < 1/(1+r)$. Then the marginal rate of substitution of x_1 for x_2 according to the welfare utility u, which is given by $\mathrm{MRS}_u = 1/\delta$, is strictly larger than the absolute value of the slope of the budget line, which is given by $1+r$. The unique u-optimal element from any compact $S \subseteq X(r, y)$ is thus the unique alternative that maximizes x_1 in S. From $\mathrm{MRS}_{u_S} = 1/\delta_S = \gamma/\delta \geq 1/\delta$ it follows that this is also the unique u_S-optimal element in S. Hence each u_S-optimal element in S is weakly u-better than each u_L-optimal element. If instead $\delta = 1/(1+\overline{r})$ and $r = \overline{r}$ holds,

we have $\text{MRS}_u = 1 + r$ and all elements in any $S \subseteq X(r, y)$ are u-optimal. It again follows that each u_S-optimal element in S is weakly u-better than each u_L-optimal element. Thus f_S is a weakly successful nudge over f_L and hence an optimal nudge.

Case 2: $1/(1+\underline{r}) \leq \delta$. Analogous arguments imply that f_L is an optimal nudge in that case.

Case 3: $1/(1+\overline{r}) < \delta < 1/(1+\underline{r})$. If $\gamma = 1$, the utility functions u, u_S, and u_L all coincide, which implies that each frame is a weakly successful nudge over the other, and hence none of them is dominated. Then assume $\gamma > 1$. Choose $(r, y) \in C$ such that $\delta/\gamma < 1/(1+r) < \delta$, which exists because C is connected. Consider $S = X(r, y)$. From $\text{MRS}_{u_L} < \text{MRS}_u < 1 + r$ it then follows that $(0, y(1 + r))$ is the unique u-optimal element in S and also the unique u_L-optimal element in S. By contrast, from $\text{MRS}_{u_S} > 1 + r$ it follows that $(y, 0)$ is the unique u_S-optimal element in S. Hence the u_S-optimal element is not weakly u-better than the u_L-optimal element, and therefore f_S is not a weakly successful nudge over f_L. Analogous arguments for some $(r, y) \in C$ with $\delta < 1/(1 + r) < \gamma\delta$ imply that f_L is also not a weakly successful nudge over f_S. Hence none of the two frames is dominated.

B.2 Additional Material

B.2.1 Complexities for the Strong Priming Model

Expected complexity, geometric distribution. Fix some $\rho \in (0, 1)$ and let

$$p_i = \rho^{i-1} \left(\frac{1 - \rho}{1 - \rho^m} \right)$$

for each $i = 1, 2, \ldots, m$, where $m = m_X!$ for convenience. Note that this is indeed a probability distribution, because $p_i \in (0, 1)$ and

$$\sum_{i=1}^{m} p_i = \left(\frac{1 - \rho}{1 - \rho^m} \right) \sum_{i=1}^{m} \rho^{i-1} = \left(\frac{1 - \rho}{1 - \rho^m} \right) \left(\frac{1 - \rho^m}{1 - \rho} \right) = 1,$$

where the second equality follows from a standard result about the geometric sequence. The expression for $\bar{n}$ in Proposition 2.9 can then be written as

$$\bar{n} = \left(\frac{1 - \rho}{1 - \rho^m} \right) \sum_{i=1}^{m-1} \rho^{i-1} i + \left(\frac{1 - \rho}{1 - \rho^m} \right) \rho^{m-1} (m - 1).$$

Using the standard result that

$$\sum_{i=1}^{m-1} \rho^{i-1} i = \left(\frac{1 - \rho^m}{(1 - \rho)^2} \right) - \left(\frac{m\rho^{m-1}}{1 - \rho} \right),$$

we can further simplify to

$$\bar{n} = \left(\frac{1}{1 - \rho} \right) + \left(\frac{(1 - \rho)(m - 1)\rho^{m-1} - m\rho^{m-1}}{1 - \rho^m} \right).$$

Due to $\rho \in (0, 1)$, the second term vanishes as $m \to \infty$. Hence $\lim_{m_X \to \infty} \bar{n} = 1/(1 - \rho)$.

Expected complexity, uniform distribution. Let $p_i = 1/m$ for each $i = 1, 2, \ldots, m$. The expression for $\bar{n}$ in Proposition 2.9 can then be written as

$$\bar{n} = \frac{1}{m}\sum_{i=1}^{m-1} i + \left(\frac{m-1}{m}\right) = \left(\frac{m-1}{2}\right) + \left(\frac{m-1}{m}\right) = (m_X! - 1)\left(\frac{1}{2} + \frac{1}{m_X!}\right),$$

which is of the same order of magnitude as the previously given $n = m_X! - 1$.

Generalized complexity, geometric distribution. For the geometric distribution, the LHS of the inequality in Proposition 2.10 can be rewritten as

$$\sum_{j=1+k}^{m-1} p_{j+1} = \left(\frac{1-\rho}{1-\rho^m}\right)\sum_{j=1+k}^{m-1}\rho^j = \left(\frac{1-\rho}{1-\rho^m}\right)\left(\frac{\rho^{1+k}-\rho^m}{1-\rho}\right) = \frac{\rho^{k+1}-\rho^m}{1-\rho^m}.$$

Thus $n(q)$ is the smallest integer $k \geq 0$ for which

$$\frac{\rho^{k+1}-\rho^m}{1-\rho^m} \leq \left(\frac{1-\rho}{1-\rho^m}\right)\left(\frac{1-q}{q}\right),$$

or

$$\rho^k \leq \rho^{m-1} + \left(\frac{1-\rho}{\rho}\right)\left(\frac{1-q}{q}\right).$$

For $q = 1$ this implies $n(1) = m-1$. Since the RHS of the inequality converges to $((1-\rho)/\rho)((1-q)/q)$ as $m \to \infty$, for $q < 1$ we obtain that $n(q)$ must converge to the smallest integer $k \geq 0$ for which

$$\rho^k \leq \left(\frac{1-\rho}{\rho}\right)\left(\frac{1-q}{q}\right)$$

holds. Hence

$$\lim_{m_X\to\infty} n(q) = \max\left\{\left\lceil \frac{\log\left(\frac{1-\rho}{\rho}\frac{1-q}{q}\right)}{\log\rho}\right\rceil, 0\right\}.$$

Generalized complexity, uniform distribution. For the uniform distribution, the condition in Proposition 2.10 becomes that $n(q)$ is the smallest integer $k \geq 0$ for which

$$k \geq (m-1) - \left(\frac{1-q}{q}\right)$$

holds. Hence we obtain

$$n(q) = \max\left\{\left\lceil (m-1) - \left(\frac{1-q}{q}\right)\right\rceil, 0\right\}.$$

B.2.2 Experimental Instructions

The following contains screenshots of the instructions and the experiment itself. Subjects were recruited on Amazon Mechanical Turk, where we restricted eligibility to participate to US subjects with an experience of at least 500 approved MTurk HITs and an approval rate of at least 95%. Figure B.1 shows the description of the HIT on MTurk. Upon participation subjects were taken to an external website where we set up the experiment using Qualtrics. Subjects first had to accept the consent form in Figure B.2, which all of the 1064 subjects did. We then collected

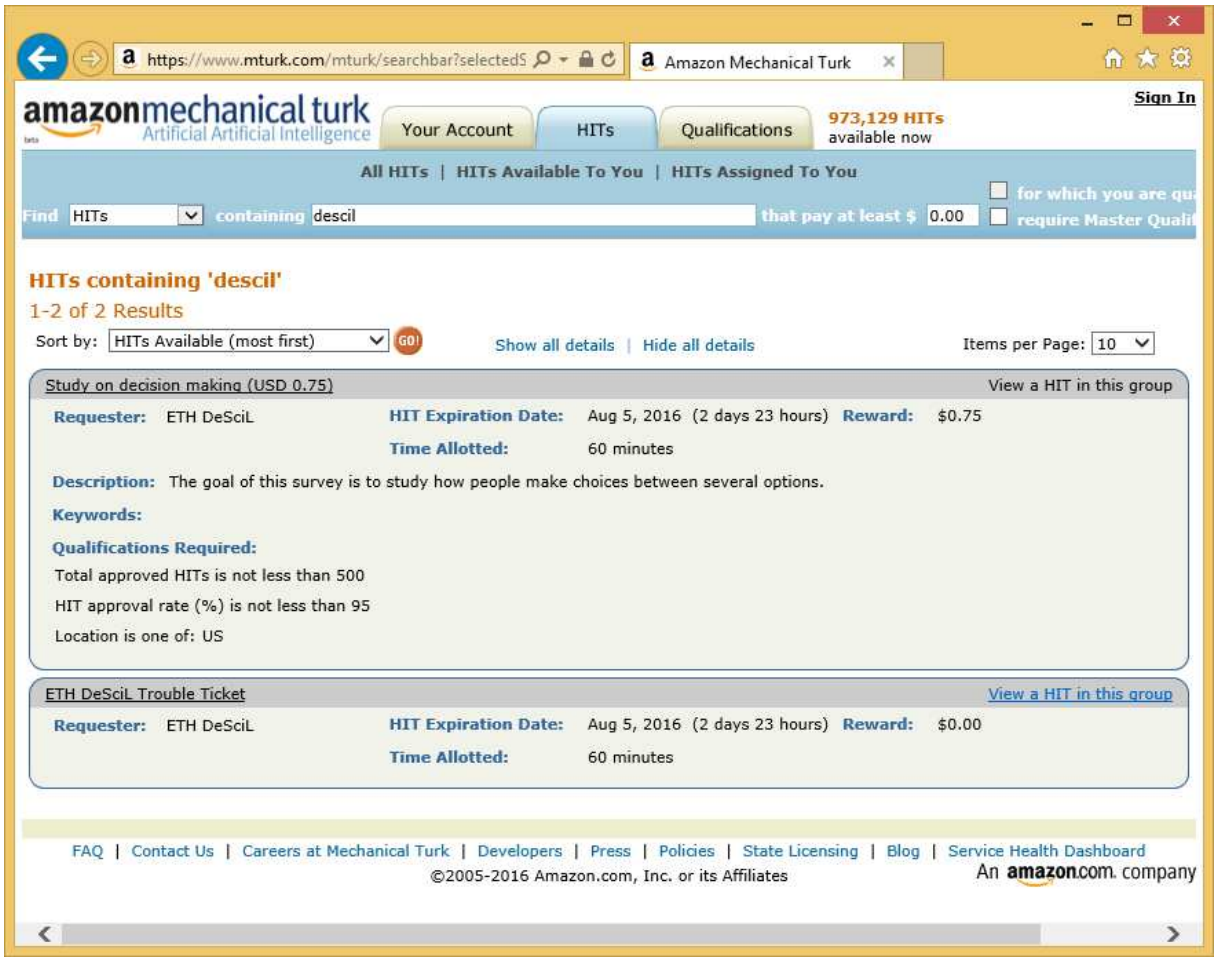


Figure B.1: Screenshot of the description of the HIT on MTurk.

some demographics (Figure B.3), before subjects faced two pairs of questions about intertemporal choice as shown in Figures B.4 and B.5. The value of the gift card was randomly assigned (either \$75 or \$85) and the order of the questions was randomized, too. Finally, subjects were given an exit code with which they could demand payment of MTurk.

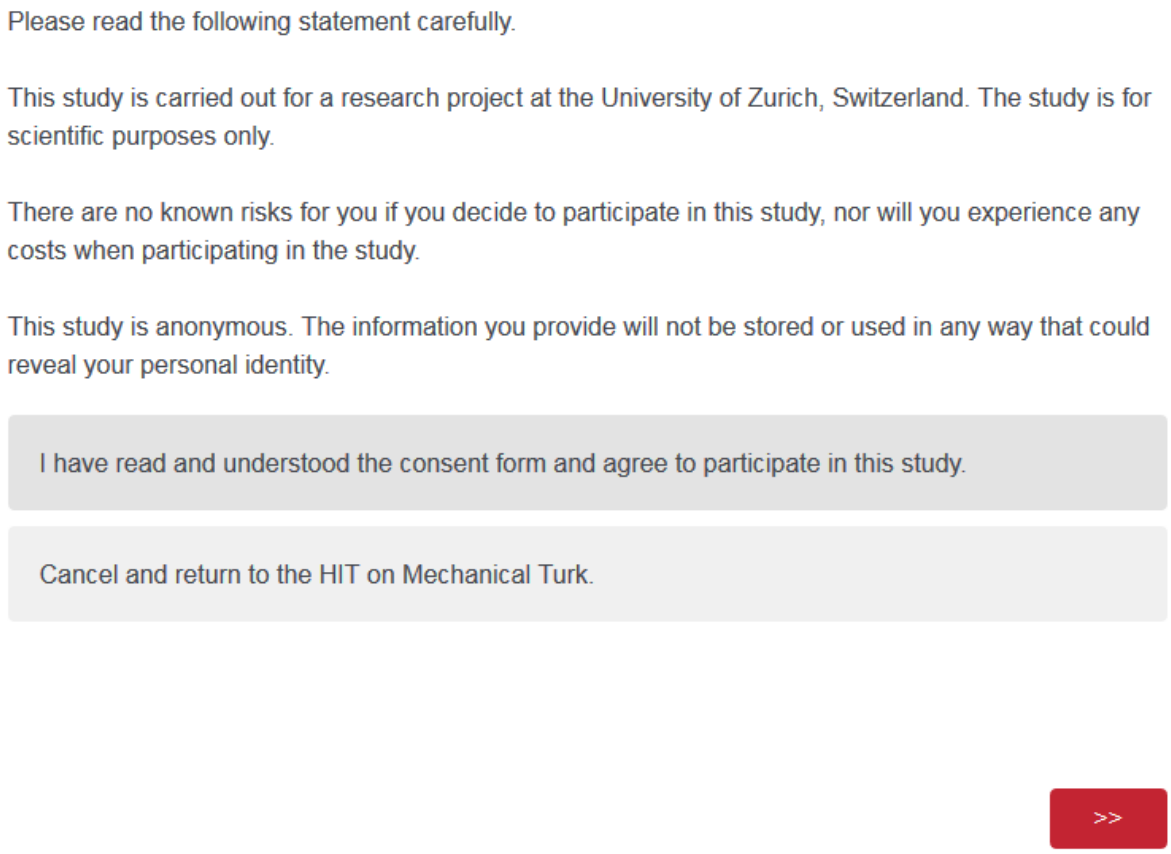


Figure B.2: Screenshot of the first page of the experiment.

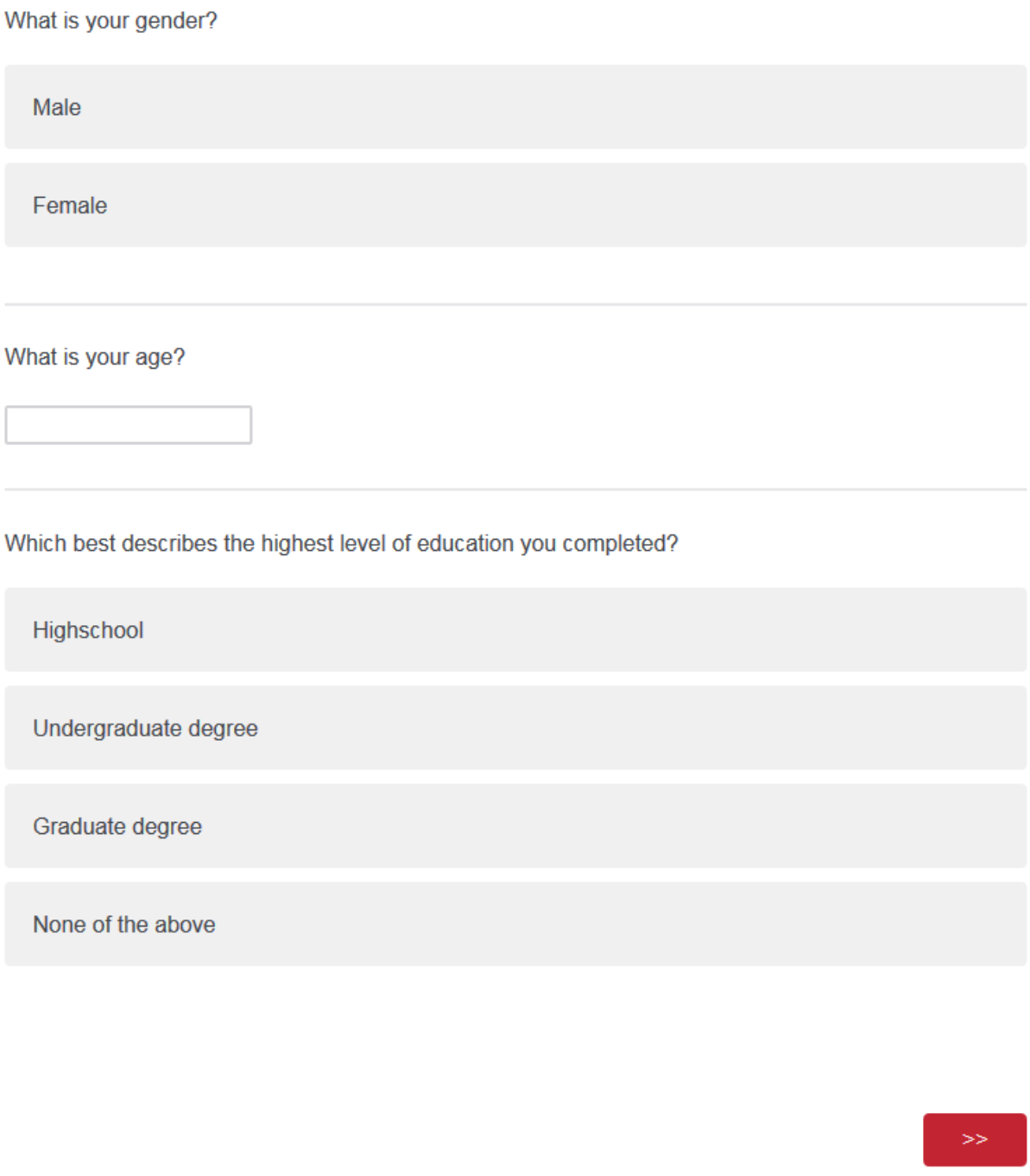


Figure B.3: Screenshot of the second page of the experiment.

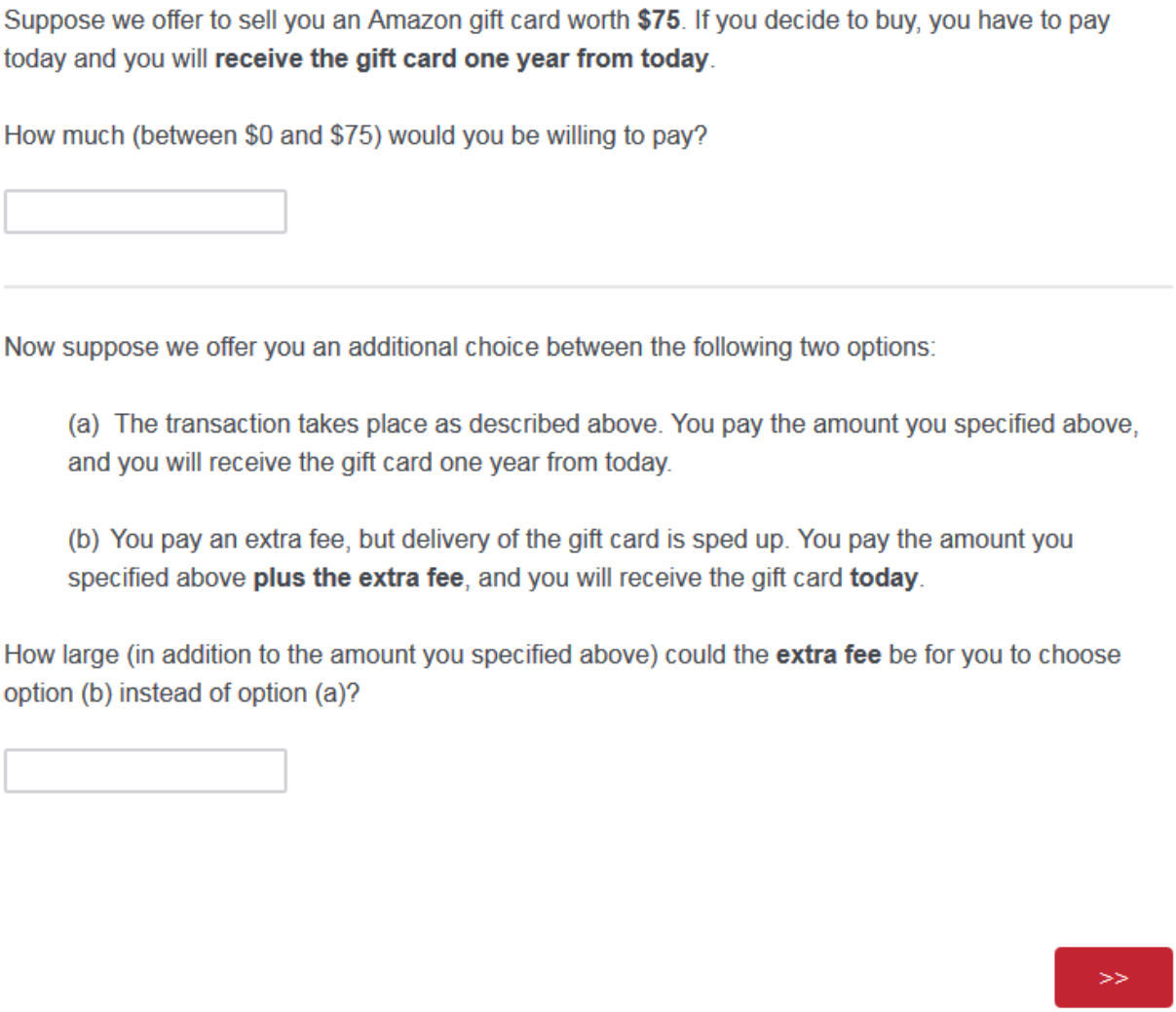


Figure B.4: Screenshot of the third page of the experiment.

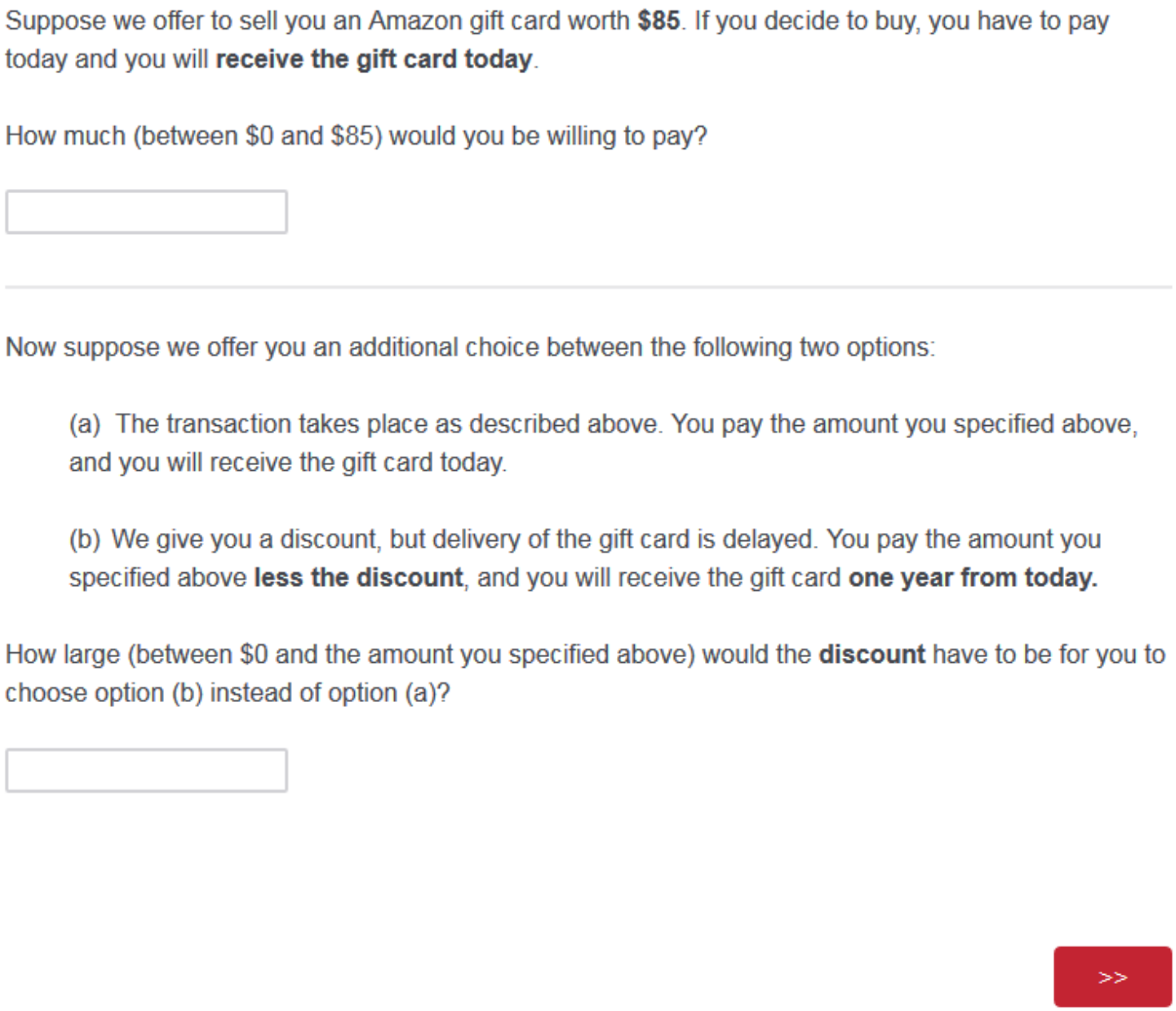


Figure B.5: Screenshot of the fourth page of the experiment.

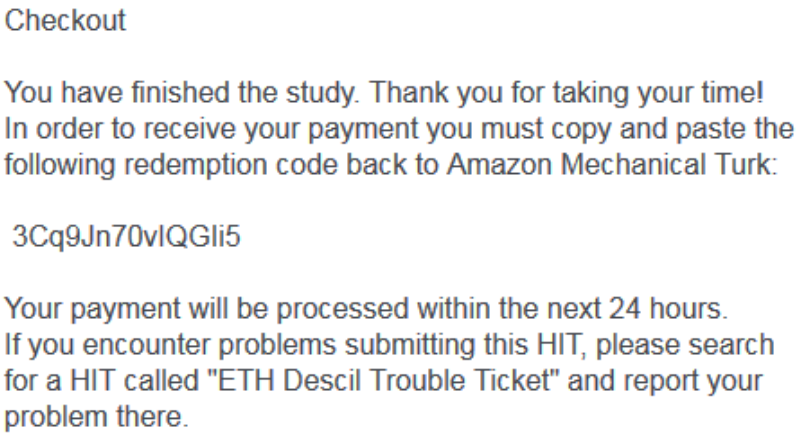


Figure B.6: Screenshot of the fifth and final page of the experiment.

C Appendix: Chapter 3

C.1 Proofs

C.1.1 Proof of Proposition 3.1

The Game

The contest $\Gamma = \langle E, p, m, n, T \rangle$ induces the following extensive form game of incomplete information: $\mathcal{G} = \langle I, H, \alpha, F, (\mathcal{I}_i)_{i \in I}, (u_i)_{i \in I} \rangle$. The set of players is $I = \{0, 1, \ldots, n\}$, where player 0 is the principal and players 1 to n are the agents. The set H is the set of histories, where the set of terminal histories is denoted Z and the actions available after the non-terminal history h is denoted $A(h) = \{a : (h, a) \in H\}$. The function α assigns to each non-terminal history a subset of I, i.e., α is the player function. The set of initial histories is the set of the states of the world. The true initial history is $\theta \in \Theta^{NT}$, where each element $\theta_{it} \in \Theta$ (where $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$) is drawn i.i.d. from the probability distribution F. An agent i who conducts research in period t receives value equal to θ_{it}. For each player $i \in I$ a partition $\mathcal{I}_i$ of $\{h \in H : \alpha(h) = i\}$ with the property that $A(h) = A(h')$, whenever h and h' are in the same member of the partition. The function $u_i : Z \to \mathbb{R}$ maps for each player i the payoff at each of the terminal nodes. For the agents the payoffs are determined by the research costs they have incurred, the entry fee they pay (or receive) if they enter the contest, and the transfers they receive. The principal's payoff is determined by the value of the innovation she gets, the entry fees of the participants, and the transfers she makes to the agents. In what follows we will use the terms doing research and investing (in research) interchangeably.

Timing

Period 0**:**

- The principal announces the contest $\Gamma = \langle E, p, m, n, T \rangle$.
- All agents decide whether to enter or not. If more than n agents want to participate, n are selected at random. All agents who enter the contest pay the entry fee E.

Period $t < T$**:**

- Stage 1: Each agent simultaneously decides whether to perform research at cost C. Agents do not observe the actions taken by their competitors.
- Stage 2: Each agent i who conducted research receives value equal to θ_{it}. All other agents receive value 0.

- Stage 3: Having privately observed the value of their innovation, agents simultaneously decide whether to privately submit their best innovation.

- Stage 4: The principal observes the set of submissions (if any) and decides whether to declare a winner. If a winner is declared the contest stops, the principal obtains the highest available innovation and the agent who submitted the winning innovation receives the prize p. If the contest continues the agent randomly pays the progress prize m to an agent.

Period T (if $T < \infty$):

- Stages 1-3: As above.

- Stage 4: The contest stops and the principal has to declare a winner and pay the prize p to the winner.

Equilibrium Candidate

Denote with $\theta^i|h$ the highest value available to player i at history h. For agents, this is the highest value they have so far discovered. For principal, this is the highest value currently submitted. The equilibrium candidate (σ, μ) is defined as follows:

Agents If $A(h) = \{\text{Invest, Not Invest}\} = \{I, NI\}$ then

$$\sigma^i(h) = \begin{cases} I & \text{if } \theta^i|h < \theta^K \\ NI & \text{else} \end{cases}. \tag{C.1}$$

If $A(h) = \{\text{Submit, Not Submit}\} = \{S, NS\}$ then

$$\sigma^i(h) = \begin{cases} S & \text{if } \theta^i|h \geq \theta^g \\ NS & \text{else} \end{cases}. \tag{C.2}$$

Equilibrium beliefs of agent i are as follows. Denote a history in the period t' as $h_{t'}$. Let the last period when the agent i has not observed a deviation by the principal be $t^e|h'_t$. This means that in period t^e, agent i did not submit and that in all periods $t^e+1, \ldots, t'-1$ the agent i submitted a value over θ^g but the principal did not end the contest. The beliefs that the element of state of the world $\tilde{\theta}$ in some period t and for a player j, where the true state of the world is θ_{jt} are given by the following cases.

- Own elements of the state of the world $(i = j)$:

$$\mu^i_{jt}(\theta^k|h_{t'}) = \begin{cases} 1 & \text{if } t \leq t',\ a_{it}|h_{t'} = I \text{ and } \theta^k = \theta_{jt} \\ 0 & \text{if } t \leq t',\ a_{it}|h_{t'} = I \text{ and } \theta^k \neq \theta_{jt} \\ F(\theta^k) - F(\theta^{k-1}) & \text{else} \end{cases}. \tag{C.3}$$

- Others' elements of the state of the world ($i \neq j$):

$$\mu^i_{jt}(\theta^k|h_{t'}) = \begin{cases} \frac{F(\theta^k)-F(\theta^{k-1})}{F(\theta^{g-1})} & \text{if } t \leq t^e|h_{t'} \text{ and } \theta^k < \theta^g \\ 0 & \text{if } t \leq t^e|h_{t'} \text{ and } \theta^k \geq \theta^g \\ F(\theta^k) - F(\theta^{k-1}) & \text{else} \end{cases}. \tag{C.4}$$

For own elements, the agent learns exact state if he invests, if he does not, or if the chance to invest has not occurred yet, he holds initial beliefs. For the others' elements, once a period starts after a no deviation from the principal, the agent concludes that everybody has invested up to that point and that nobody has a value higher than θ^g. This implies that each individual θ_{jt} is drawn from the truncated distribution. If the agent observes that the principal deviated, he learns nothing about the realization of the state in that period, hence he should hold the initial beliefs. For all the states which have not been revealed yet, the agent holds initial beliefs.

Principal Consider any information set at which the principal is moving. The principal stops the game if and only if there has been a submission of value at least θ^g. For any agent who has not submitted an innovation, the principal believes that research has been conducted in every period, yet the draws were always below θ^g. For a firm which submitted an innovation, the principal believes that research has been conducted in every period and that the submission is the currently highest value.

Proof

The proof proceeds as follows. First, three lemmas show that no profitable one-shot deviation exists after any history of the game. Second, we prove that no one-shot deviation implies that no profitable deviation exists.

Lemma C.1 *There exists no profitable one-shot deviation for the principal.*

Proof. If $T < \infty$ there is a final period. In this final period the principal has to declare some agent the winner and pay the prize p. Thus, consider any period $t < T$. Suppose an innovation of value $\theta^k \geq \theta^g$ has been submitted to the principal. Stopping yields $\theta^k - p$, whereas continuing yields $-m + \delta(\Delta(\theta^k, n) - p)$. Thus, stopping is optimal whenever

$$m \geq p(1-\delta) + \delta\Delta(\theta^k, n) - \theta^k.$$

Recall that $m = p(1-\delta) + \delta\Delta(\theta^g, n) - \theta^g$. Simple algebra shows that $\Delta(\theta, n) - \theta$ is strictly decreasing in θ. Thus, the principal will stop the contest if a value $\theta^k \geq \theta^g$ has been submitted. Suppose now a value $\theta^k < \theta^g$ has been submitted. We will show in three steps that stopping is not optimal.

Step 1. Denote with $V_t(\mu, \pi|\theta^k)$ the value to the principal of having the highest value θ^k in period t given that she follows the equilibrium candidate. Then, it follows that $V_t(\mu, \pi|\theta^g) = \theta^g - p$. In

this step, we will show that also

$$V_t(\mu,\pi|\theta^g) = -m + \delta\left[F(\theta^g)^n V_{t+1}(\mu,\pi|\theta^g) + \sum_{j=g+1}^{K}(F(\theta^j)^n - F(\theta^{j-1})^n)V_{t+1}(\mu,\pi|\theta^j)\right]$$

for any t. It suffices to show that

$$\theta^g - p = -m + \delta\left[F(\theta^g)^n V_{t+1}(\mu,\pi|\theta^g) + \sum_{j=g+1}^{K}(F(\theta^j)^n - F(\theta^{j-1})^n)V_{t+1}(\mu,\pi|\theta^j)\right]$$

Obviously, for any $\theta^j \geq \theta^g$ it holds $V_{t+1}(\mu,\pi|\theta^j) = \theta^j - p$. Substituting and rearranging we obtain

$$m = p(1-\delta) + \delta\Delta(\theta^g, n) - \theta^g$$

which holds by definition for any t.

Step 2. In this step we show that for any pair θ^k, θ^{k+1} such that $\theta^{k+1} \leq \theta^g$ it holds that

$$V_t(\mu,\pi|\theta^{k+1}) - V_t(\mu,\pi|\theta^k) = \delta F^n(\theta^k)(V_{t+1}(\mu,\pi|\theta^{k+1}) - V_{t+1}(\mu,\pi|\theta^k)).$$

We have

$$V_t(\mu,\pi|\theta^k) = -m + \delta\left(F^n(\theta^k)V_{t+1}(\mu,\pi|\theta^k) + \sum_{j=k+1}^{K}(F^n(\theta^j) - F^n(\theta^{j-1}))V_{t+1}(\mu,\pi|\theta^j)\right)$$

and an equivalent expression holds for $V_t(\mu,\pi|\theta^{k+1})$. Expanding $V_t(\mu,\pi|\theta^{k+1}) - V_t(\mu,\pi|\theta^k)$ we get:

$$\begin{aligned}
&V_t(\mu,\pi|\theta^{k+1}) - V_t(\mu,\pi|\theta^k)\\
&= \delta\left(F^n(\theta^{k+1})V_{t+1}(\mu,\pi|\theta^{k+1}) - F^n(\theta^k)V_{t+1}(\mu,\pi|\theta^k) - (F^n(\theta^{k+1}) - F^n(\theta^k))V_{t+1}(\mu,\pi|\theta^{k+1})\right)\\
&= \delta F^n(\theta^k)(V_{t+1}(\mu,\pi|\theta^{k+1}) - V_{t+1}(\mu,\pi|\theta^k)).
\end{aligned}$$

Step 3. In this step we show that for any $\theta^k < \theta^g$, the principal does not stop the contest, i.e., there is no profitable one-shot deviation. Stopping the contest in period t yields the payoff of $\theta^k - p$. Thus, it suffices to show that $V_t(\mu,\pi|\theta^k) > \theta^k - p$ for any $\theta^k < \theta^g$. Observe that $V_t(\mu,\pi|\theta^g) - \theta^g + p = 0$. We will show that $V_t(\mu,\pi|\theta^k) - \theta^k$ is strictly decreasing in θ^k for any $\theta^k < \theta^g$. The result then follows. This is equivalent to

$$V_t(\mu,\pi|\theta^{k+1}) - V_t(\mu,\pi|\theta^k) < \theta^{k+1} - \theta^k,$$

where $\theta^{k+1} \leq \theta^g$.

Suppose $T < \infty$. Then, we have $V_T(\mu,\pi|\theta^{k+1}) - V_T(\mu,\pi|\theta^k) = \theta^{k+1} - \theta^k$ because the contest

ends in that period. Iterating Step 2 it follows that

$$\begin{aligned} V_t(\mu,\pi|\theta^{k+1}) - V_t(\mu,\pi|\theta^k) &< V_T(\mu,\pi|\theta^{k+1}) - V_T(\mu,\pi|\theta^k) \\ &= \theta^{k+1} - \theta^k \end{aligned}$$

and thus a one-shot deviation is not profitable.

Suppose $T = \infty$. Then we have $V_t(\mu,\pi|\theta^k) = V(\mu,\pi|\theta^k)$ for all t and all θ^k. Moreover, for any $\theta^k < \theta^g$ we have $V(\mu,\pi|\theta^k) = V(\mu,\pi|\theta^{k+1})$. Further,

$$\begin{aligned} V(\mu,\pi|\theta^{g-1}) &= F^n(\theta^{g-1})\delta V(\mu,\pi|\theta^{g-1}) + \delta\left(\sum_{j=g}^{K}(F^n(\theta^j) - F^n(\theta^{j-1}))\theta^j\right) - m \\ &= \frac{\delta}{1-F^n(\theta^{g-1})\delta}\left(\sum_{j=g}^{K}(F^n(\theta^j) - F^n(\theta^{j-1}))\theta^j\right) - \frac{m}{1-F^n(\theta^{g-1})\delta}. \end{aligned}$$

We will now show that the principal will not want to deviate with value θ^{g-1}, that is, we show that $V(\mu,\pi|\theta^{g-1}) > \theta^{g-1} - p$. Rewriting this inequality and replacing m we obtain

$$\frac{\delta}{1-F^n(\theta^{g-1})\delta}\left(\Delta(\theta^g,n) - F^n(\theta^{g-1})\theta^g\right) - \frac{p(1-\delta)+\delta\Delta(\theta^g,n)-\theta^g}{1-F^n(\theta^{g-1})\delta} > \theta^{g-1} - p.$$

Collecting terms and further simplifying we obtain

$$\begin{aligned} -\frac{\delta}{1-F^n(\theta^{g-1})\delta}F^n(\theta^{g-1})\theta^g - \frac{p(1-\delta)-\theta^g}{1-F^n(\theta^{g-1})\delta} &> \theta^{g-1} - p \\ \theta^g(1-\delta F^n(\theta^{g-1})) - p(1-\delta) &> (\theta^{g-1}-p)(1-F^n(\theta^{g-1})\delta) \\ (\theta^g - \theta^{g-1})(1-\delta F^n(\theta^{g-1})) &> -p\delta(1-F^n(\theta^{g-1})) \end{aligned}$$

which always holds.

Thus, whenever the principal has obtained a value of θ^{g-1} she will not stop the contest. Further, for any $\theta^k, \theta^{k+1} < \theta^g$ we have

$$V(\mu,\pi|\theta^{k+1}) - \theta^{k+1} - (V_t(\mu,\pi|\theta^k) - \theta^k) = \theta^k - \theta^{k+1} < 0.$$

implying that she will also not stop it for any other $\theta^k < \theta^g$. ■

Lemma C.2 *There is exists profitable one-shot deviation at the submission stage for the agent.*

Proof. Observe that submitting an innovation that has value below θ^g is never profitable. Thus we only need to consider the decision of an agent who has an innovation of value $\theta^k \geq \theta^g$. Suppose the state of the world is such that another agent has a value $\theta \geq \theta^k$. Then, submitting yields a weakly higher payoff, as it could mean the agent wins the prize, whereas not submitting yields zero as the contest ends for sure. Finally, suppose the state of the world is such that no other agent has a value $\theta \geq \theta^k$.

We need to consider two cases: when $\theta^k = \theta^K$ and when $\theta^k < \theta^K$. Suppose first that $\theta^k = \theta^K$. The payoff of following the equilibrium strategy and submitting is p. One-shot deviation is to

not submit, then not do research, and then submit. The payoff of this deviation is

$$\frac{m}{n} + \delta \mathcal{P}_{t+1}(\pi' | \theta^K, t)p$$

where $\mathcal{P}_{t+1}(\pi' | \theta^K, t)$ is the probability that the agent wins the contest in period $t+1$ given that he has the quality θ^K in period t and follows the deviation strategy π'. The deviation will not be profitable if

$$p \geq \frac{m}{n} + \delta \mathcal{P}_{t+1}(\pi' | \theta^K, t)p.$$

Since $m = (1-\delta)p + \delta\Delta(\theta^g, n) - \theta^g$ and $\mathcal{P}_{t+1}(\pi' | \theta^K, t) < 1$, for all p large enough it holds

$$\frac{(1-\delta)p}{n} + \delta p \geq \frac{m}{n} + \delta \mathcal{P}_{t+1}(\pi' | \theta^K, t)p.$$

Observe that

$$\begin{aligned} \frac{(1-\delta)p}{n} + \delta p &= (1 + (n-1)\delta)\frac{p}{n} \\ &\leq (1 + (n-1))\frac{p}{n} \\ &\leq p \end{aligned}$$

so that for p large enough the deviation will not be profitable.

In the case $\theta^k < \theta^K$ the one-shot deviation is to not submit, invest, and then submit. However, observe that the quality in the next period cannot be greater than θ^K, which implies that the deviation payoff is less than $m/n + \delta(\mathcal{P}_{t+1}(\pi | \theta^K, t+1)p - C)$. As this is less than in the previous case, this deviation is also not profitable. ■

Lemma C.3 *There exists no profitable one-shot deviation at the research stage for the agent.*

Proof. Suppose that the highest quality agent i has in period t is θ^k. In what follows, we show that for a sufficiently high p investing is optimal for all $\theta^k < \theta^K$ and it is not optimal for $\theta^k = \theta^K$. Let π' be a strategy profile that coincides with the equilibrium candidate π with the exception of the agent i's action in the investment stage in period t. Thus, it is a one-shot deviation. First note that a deviation in the case $\theta^k = \theta^K$ would imply investing when the agent has the highest feasible quality. This is trivially never optimal, as the agent incurs research costs without an increase in quality. Thus, focus on the case $\theta^k < \theta^K$ where a deviation is to not invest.

Denote the expected utility of agent i following the strategy π from period t in which his highest quality is θ^k with $U_i(\pi | \theta^k, t)$. A one-shot deviation is not profitable if

$$U_i(\pi | \theta^k, t) - U_i(\pi' | \theta^k, t) \geq 0. \tag{C.5}$$

As before, let $\mathcal{P}_s(\pi | \theta^k, t)$ be the probability that the agent i wins the contest in period $s \geq t$, again following the strategy π from period t in which the highest quality was θ^k. In period t,

when the deviation takes place, we have

$$\mathcal{P}_t(\pi|\theta^k,t) > \mathcal{P}_t(\pi'|\theta^k,t),$$

as the investment strictly increases the probability of winning. Further, for all $t < s < T$ (i.e., any subsequent period after the deviation except the last one) we have

$$\mathcal{P}_s(\pi|\theta^k,t) = \mathcal{P}_s(\pi'|\theta^k,t),$$

and in the final period (if there is any) we have

$$\mathcal{P}_T(\pi|\theta^k,t) > \mathcal{P}_T(\pi'|\theta^k,t).$$

Now consider the case $\theta^g \leq \theta^k < \theta^K$. In this case, the game will end with certainty in period t and the LHS of inequality (C.5) reads

$$U_i(\pi|\theta^k,t) - U_i(\pi'|\theta^k,t) = -C + p(\mathcal{P}_t(\pi|\theta^k,t) - \mathcal{P}_t(\pi'|\theta^k,t)).$$

Thus, for a sufficiently high p, not investing is not a profitable deviation.

The only remaining case is $\theta^k < \theta^g$, which we now consider. Observe that we can write the expected utilities of the strategies π and π' in the following way

$$U_i(\pi|\theta^k,t) = p\mathcal{P}_t(\pi|\theta^k,t) + \sum_{s=t+1}^{T}\left(\mathcal{P}_s(\pi|\theta^k,t)p + \frac{m}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{s-t} - K$$
$$U_i(\pi'|\theta^k,t) = \delta F^{n-1}(\theta^{g-1})\sum_{s=t+1}^{T}\left(\mathcal{P}_s(\pi'|\theta^k,t)p + \frac{m}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{s-t-1} - K'$$

where K and K' collect all cost terms. Furthermore, note that $m = (1-\delta)p + \delta\Delta(\theta^g,n) - \theta^g$, so if we redefine K and K' to contain all terms not containing p we can write

$$U_i(\pi|\theta^k,t) = p\mathcal{P}_t(\pi|\theta^k,t) + \sum_{s=t+1}^{T}\left(\mathcal{P}_s(\pi|\theta^k,t)p + \frac{(1-\delta)p}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{s-t} - K$$
$$U_i(\pi'|\theta^k,t) = \delta F^{n-1}(\theta^{g-1})\sum_{s=t+1}^{T}\left(\mathcal{P}_s(\pi'|\theta^k,t)p + \frac{(1-\delta)p}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{s-t-1} - K'.$$

Then, for sufficiently high p, the inequality C.5 will be satisfied if

$$p\mathcal{P}_t(\pi|\theta^k,t) + \sum_{s=t+1}^{T}\left(\mathcal{P}_s(\pi|\theta^k,t)p + \frac{(1-\delta)p}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{s-t} \geq$$
$$\delta F^{n-1}(\theta^{g-1})\sum_{s=t+1}^{T}\left(\mathcal{P}_s(\pi'|\theta^k,t)p + \frac{(1-\delta)p}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{s-t-1}.$$

Simplifying and collecting terms

$$\mathcal{P}_t(\pi|\theta^k,t) \geq \delta F^{n-1}(\theta^{g-1})\times$$
$$\sum_{s=t+1}^{T}\left(\mathcal{P}_s(\pi'|\theta^k,t)+\frac{(1-\delta)}{n\delta}-F(\theta^{g-1})\left(\mathcal{P}_s(\pi|\theta^k,t)+\frac{(1-\delta)}{n\delta}\right)\right)(\delta F^n(\theta^{g-1}))^{s-t-1}.$$

Since $\mathcal{P}_s(\pi|\theta^k,t) \geq \mathcal{P}_s(\pi'|\theta^k,t)$ for all s, a sufficient condition for the above inequality to hold is

$$\mathcal{P}_t(\pi|\theta^k,t) \geq \delta F^{n-1}(\theta^{g-1})(1-F(\theta^{g-1}))\times$$
$$\sum_{s=t+1}^{T}\left(\mathcal{P}_s(\pi'|\theta^k,t)+\frac{(1-\delta)}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{s-t-1}$$

Furthermore, $\mathcal{P}_t(\pi|\theta^k,t) \geq F^{n-1}(\theta^{g-1})(1-F(\theta^{g-1}))$ as the probability of winning certainly includes all states of the world where the agent i obtains the quality above θ^{g-1} while no one else does. Thus, a sufficient condition for the above inequality to hold is

$$1 \geq \delta\sum_{s=t+1}^{T}\left(\mathcal{P}_s(\pi'|\theta^k,t)+\frac{(1-\delta)}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{s-t-1}$$

which we can rewrite as

$$1 \geq \delta\sum_{s=t+1}^{T-1}\left(\mathcal{P}_s(\pi'|\theta^k,t)+\frac{(1-\delta)}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{s-t-1}$$
$$+\delta\left(\mathcal{P}_T(\pi'|\theta^k,t)+\frac{(1-\delta)}{n\delta}\right)(\delta F^n(\theta^{g-1}))^{T-t-1}.$$

Observe that $\mathcal{P}_s(\pi'|\theta^k,t) = (1-F^n(\theta^{g-1}))/n$ since $1-F^n(\theta^{g-1})$ is the probability that the contest ends and by symmetry the agents have equal chances of winning it. Furthermore, $\mathcal{P}_s(\pi'|\theta^k,t) < 1$. Hence, a sufficient condition for the above inequality to hold is

$$1 \geq \frac{1-\delta F^n(\theta^{g-1})}{n}\frac{1-\delta F^n(\theta^{g-1})^{T-t-1}}{1-\delta F^n(\theta^{g-1})}+\frac{(1+(n-1)\delta)}{n}(\delta F^n(\theta^{g-1}))^{T-t-1}$$
$$\geq \frac{1}{n}(1-\delta F^n(\theta^{g-1})^{T-t-1})+\frac{(1+(n-1)\delta)}{n}(\delta F^n(\theta^{g-1}))^{T-t-1}$$

which holds for T finite since both $1/n$ and $(1+(n-1)\delta)/n$ are smaller than one, and the RHS equals $1/n$ for infinite T. ■

We conclude the proof by showing that since no one-shot deviation exists, then no profitable deviation exists at all. First, observe that if T is finite, then the result follows by Theorem 1 of Hendon, Jacobsen, and Sloth (1996). If T is infinite and $\delta < 1$, then the game is continuous at infinity and the result follows by Corollary 2 of Hendon et al. (1996). The only remaining case is T infinite and $\delta = 1$. Our proof is an adaptation of the standard argument (e.g. the proof of Corollary 2 in Hendon et al. (1996) and the proof of Theorem 4.2 in Fudenberg and Tirole (1991)). As before, denote with (σ,μ) the candidate equilibrium. From Lemmas C.1-C.3 we know that no profitable OSD exists. Suppose that some other profitable deviation exists. We consider deviations by agents and by the principal separately.

First we consider deviations by agents. Then, there exists an agent i, a history h_t and a strategy $\tilde{\sigma}^i$ such that $U_i(\tilde{\sigma}^i, \sigma{-}i|h_t) > U_i(\sigma^i, \sigma{-}i|h_t)$. Let the difference be $2\epsilon > 0$. Consider an alternative strategy of agent i, $\hat{\sigma}^i_{\hat{t}}$, where $\hat{\sigma}^i_{\hat{t}}(h^t) = \tilde{\sigma}^i(h^t)$ for all $t \leq \hat{t}$ and $\hat{\sigma}^i_{\hat{t}}(h^t) = \sigma^i(h^t)$ for all $t > \hat{t}$. That is, the strategy $\hat{\sigma}^i_{\hat{t}}$ agrees with the deviation strategy $\tilde{\sigma}^i$ in the first $\hat{t}$ rounds of the contest, while it agrees with the equilibrium candidate afterwards. Observe that, regardless of the strategy employed by agent i, the probability that the contest is still active after additional t periods is at most $(F^{n-1}(\theta^{g-1}))^{t-1} < 1$. Thus, there exists $\hat{t}$ large enough, such that $U_i(\tilde{\sigma}^i, \sigma{-}i|h_t) - U_i(\hat{\sigma}^i_{\hat{t}}, \sigma{-}i|h_t) < \epsilon$. But then $U_i(\hat{\sigma}^i_{\hat{t}}, \sigma{-}i|h_t) - U_i(\sigma^i, \sigma^{-i}|h_t) > \epsilon$. However, since $\hat{\sigma}^i_{\hat{t}}$ and σ^i differ only after a finite number of histories, this would imply by Theorem 1 of Hendon et al. (1996) that σ^i admits a profitable one-shot deviation. A contradiction.

Second, suppose that there exists a history h_t and a deviation strategy for the principal $\tilde{\sigma}^0$ such that $U_0(\tilde{\sigma}^0, \sigma^{-0}|h_t) > U_0(\sigma^0, \sigma^{-0}|h_t)$. Let $\hat{\sigma}^0$ be an alternative strategy where after any history in which quality θ^K has been submitted, the principal stops the contest. After all other histories the strategies $\tilde{\sigma}^0$ and $\hat{\sigma}^0$ coincide. Since stopping the contest is the dominant strategy after θ^K has been submitted, then $U_0(\hat{\sigma}^0, \sigma^{-0}|h_{t'}) \geq U_0(\tilde{\sigma}^0, \sigma^{-0}|h_{t'})$ for all $h_{t'}$ and in particular $U_0(\hat{\sigma}^0, \sigma^{-0}|h_t) > U_0(\sigma^0, \sigma^{-0}|h_t)$. However, observe that when following the strategy $(\hat{\sigma}^0, \sigma^{-0})$ the probability that the contest is still active after additional t periods is at most $(F^n(\theta^K))^{t-1} < 1$. But then, using the same argument as above, we can construct a strategy which outperforms σ^0 but only differs from it after a finite number of histories, reaching a contradiction again.

C.1.2 Proof of Proposition 3.2

Consider an optimal search problem with no recall, where the parameters of the search problem (C, Θ, F, δ) are as in our model. Further, like in our model, assume that in each period at most N draws from the distribution F can be made. The optimal search problem is stationary and only depends on the currently available highest outcome θ. Let $V(\theta)$ denote the value of θ. We then have that

$$V(\theta) = \begin{cases} \theta & \text{if } \theta \geq V^* \\ V^* & \text{otherwise,} \end{cases}$$

where V^* is the continuation value and is given by

$$V^* = \max_{\theta^g \in \Theta,\, n \leq N} \delta \left(-nC + F^n(\theta^{g-1})V^* + \sum_{j=g}^{K} (F^n(\theta^j) - F^n(\theta^{j-1}))\theta^j \right).$$

Observe that the domain of both θ^g and n is finite, hence a maximum exists. Denote one solution to the maximization problem with n_N^{FB} and θ_N^g. Then the optimal search policy in the case of no recall is to take n_N^{FB} draws from the distribution F in every period until a value of at least θ_N^g is obtained, and then to stop immediately. By the same arguments as in Benkert et al. (2017), the optimal search rule in the case of full recall is identical to the case of no recall. However, observe that the optimal search rule in the full-recall case is also the solution to the first-best problem in our model.

By Proposition 3.1, we know that there exists a PPC which can implement the global stopping rule θ_N^g with n_N^{FB} and $T = \infty$, thus generating the first-best surplus. Then, by setting E appropriately, the principal can extract the entire expected surplus and achieve the first-best outcome.

C.1.3 Proof of Proposition 3.3

The proposition is established in two steps. First, we provide conditions for the first-best outcome to be such that all N agents conduct research in every period until a breakthrough is achieved after which search is stopped completely. Second, we show that if Assumption 3.1 holds, then the conditions identified in the first step are satisfied.

Let $n(\theta, t)$ be a function which specifies the first-best search intensity in period t given that the highest value obtained so far is θ. Gal et al. (1981) and Morgan (1983) have shown that $n(\theta, t)$ is decreasing in θ and increasing in t. Denote with $n_\theta(\theta^k, t) = n(\theta^{k+1}, t) - n(\theta^k, t)$ and $n_t(\theta^k, t) = n(\theta^k, t+1) - n(\theta^k, t)$ the change in the optimal number of agents due to an increase in θ or t respectively.

Step 1: We begin by showing that $n(\theta^{b-1}, 1) = N$ is a sufficient condition for $n(\theta, t) = N$ for all $\theta < \theta^b$ and all $t \geq 1$. Recall that $n_t(\theta, t) \geq 0$. Thus, we have $n(\theta^{b-1}, t) = N$ for all $t \geq 1$. Further, recall that $n_\theta(\theta, t) \leq 0$. Consequently, $n(\theta^{b-1}, 1) = N$ implies $n(\theta^s, 1) = N$ for all $s \in \{1, \ldots, b-1\}$. Analogously, $n(\theta^{b-1}, T) = N$ implies $n(\theta^s, T) = N$ for all $s \in \{1, \ldots, b-1\}$. Taking this together we have $n(\theta, t) = N$ for all $\theta < \theta^b$ and all $t \geq 1$.

We next show that $n(\theta^b, T) = 0$ is a sufficient condition for $n(\theta, t) = 0$ for all $\theta \geq \theta^b$ and all $t \geq 1$. It follows from $n_t(\theta, t) \geq 0$ that $n(\theta^b, t) = 0$ for all $t \geq 1$. Further, because $n_\theta(\theta, t) \leq 0$ we must have $n(\theta, t) = 0$ for all $\theta \geq \theta^b$ and all $t \geq 1$.

Step 2: Next, we show that Assumption 3.1 implies that the conditions identified in Step 1 hold. Note that $n(\theta^b, T) = 0$ is equivalent to

$$F(\theta^b)\theta^b + \sum_{j=b+1}^{K} \left(F(\theta^j) - F(\theta^{j-1})\right)\theta^j - \theta^b < C.$$

The left-hand side of the inequality is the expected benefit of conducting research with one agent in the last period given that the an innovation of quality θ^b is already available. The right-hand side is the cost of doing research with a single agent. We can rearrange the inequality to obtain

$$\begin{aligned}\left(F(\theta^b) - 1\right)\theta^b + \left(1 - F(\theta^b)\right)\theta^b + \sum_{j=b+1}^{K} \left(F(\theta^j) - F(\theta^{j-1})\right)\left(\theta^j - \theta^b\right) < C \\ \sum_{j=b+1}^{K} \left(F(\theta^j) - F(\theta^{j-1})\right)\left(\theta^j - \theta^b\right) < C\end{aligned}$$

This inequality is satisfied for any F if

$$\left(\theta^K - \theta^b\right) \leq C$$

which holds by Assumption 3.1 for sufficiently small ε since $\Theta^K = \theta^b + r\varepsilon$.

We will now give conditions on θ^b for $n(\theta^{b-1}, 1) \geq N$. Note that to demonstrate that $n(\theta^{b-1}, 1) \geq N$ it is enough to show that having N agents conduct research is better than $N-1$. First, we cannot have more than N agents, and second, Morgan (1983, Proposition 2) shows that the expected benefit of conducting research within a period is concave in the number of agents. Thus, if it is better to have N than $N-1$ agents, it is also better than having $N-s$ for $s \geq 1$ agents. Let $V(\theta, t)$ denote the value of having quality θ in period t given an optimal continuation in subsequent periods. Then, we can write the expected payoff of having N agents conduct research in period 1 given that we have quality θ^{b-1} as

$$F^N(\theta^{b-1})V(\theta^{b-1}, 2) + \theta^b(1 - F^N(\theta^{b-1})) + M(N) - NC$$

where $M(N) = \sum_{j=b+1}^{K} \left(F^N(\theta^j) - F^N(\theta^{j-1})\right)\left(\theta^j - \theta^b\right)$. The expected payoff of having $N-1$ agents conduct research is

$$F^{N-1}(\theta^{b-1})V(\theta^{b-1}, 2) + \theta^b(1 - F^{N-1}(\theta^{b-1})) + M(N-1) - (N-1)C.$$

Thus, having N agents is better if the difference between the two inequalities is weakly positive, which reads

$$(1 - F(\theta^{b-1}))F^{N-1}(\theta^{b-1})\left[\theta^b - V(\theta^{b-1}, 2)\right] + (M(N) - M(N-1)) - C \geq 0.$$

Consider now the value θ^b such that the inequality holds with equality, i.e., such that we are indifferent between N and $N-1$ agents. This implies that we would rather have N than $N-s$ agents for $s \geq 2$ by the within-period concavity in the number of agents. Moreover, since $n_t(\theta, t) \geq 0$ the θ^b which induces indifference in period 1 is such that for any period $t \geq 2$ having N firms is weakly better than having $N-1$ firms, too. Therefore, the optimal continuation is to always employ N firms. Hence,

$$\begin{aligned} V(\theta^{b-1}, 2) =& F^{N(T-1)}(\theta^{b-1})\theta^{b-1} + (1 - F^{N(T-1)}(\theta^{b-1}))\left(\theta^b(1 - F^N(\theta^{b-1})) + M(N)\right) \\ & - NC\sum_{j=1}^{T-1} F^{Nj}(\theta^{b-1})j, \end{aligned}$$

because either we continue to have a quality of θ^{b-1} until the end, or at some point we have a breakthrough.

Therefore,

$$\begin{aligned} \theta^b \geq \bar{\theta} =& \frac{1}{1 - (1 - F^{N(T-1)}(\theta^{b-1}))(1 - F^N(\theta^{b-1}))} \times \Bigg(\frac{C - (M(N) - M(N-1))}{(1 - F(\theta^{b-1}))F^{N-1}(\theta^{b-1})} \\ & - F^{N(T-1)}(\theta^{b-1})\theta^{b-1} + (1 - F^{N(T-1)}(\theta^{b-1}))M(N) - NC\sum_{j=1}^{T-1} F^{Nj}(\theta^{b-1})j\Bigg) \end{aligned}$$

is a sufficient condition on θ^b.

C.1.4 Proof of Proposition 3.4

The result follows because the first-best is a global stopping rule with a constant number of agents (Proposition 3.3) and can thus be implemented using a PPC (Proposition 3.1).

C.1.5 Proof of Proposition 3.5

The proof proceeds in two steps. In the Step 1 we show that extending T is always beneficial for the principal in case of a PPC. In Step 2 we show by example that extending T may be harmful for the principal in case of an FPC.

Step 1: The expected research costs of implementing a global stopping threshold θ^g in a T-period PPC are given by

$$EK^g(\theta^g, N, T) = \left(\sum_{t=1}^{T-1} tF^{(t-1)N}(\theta^{g-1})(1 - F^N(\theta^{g-1})) + TF^{(T-1)N}(\theta^{g-1}) \right) NC.$$

Thus, the marginal cost of extending the contest to $T+1$ periods is

$$\begin{aligned}
&EK^g(\theta^g, N, T+1) - EK^g(\theta^g, N, T) = \\
&= \left(\sum_{t=1}^{T} tF^{(t-1)N}(\theta^{g-1})(1 - F^N(\theta^{g-1})) + (T+1)F^{TN}(\theta^{g-1}) \right) NC \\
&\quad - \left(\sum_{t=1}^{T-1} tF^{(t-1)N}(\theta^{g-1})(1 - F^N(\theta^{g-1})) + TF^{(T-1)N}(\theta^{g-1}) \right) NC \\
&= \left(-TF^{(T-1)N}(\theta^{g-1})(F^N(\theta^{g-1})) + (T+1)F^{TN}(\theta^{g-1}) \right) NC \\
&= F^{TN}(\theta^{g-1})NC
\end{aligned}$$

and the expected value of innovation is given by

$$EQ^g(\theta^g, N, T) = \sum_{k=1}^{K} \theta^k h^g(\theta^k | \theta^g, N, T),$$

where

$$h^g(\theta^k | \theta^g, N, T) = \begin{cases} F^{NT}(\theta^k) - F^{NT}(\theta^{k-1}) & k < g, \\ \sum_{t=1}^{T} F^{N(t-1)}(\theta^{g-1}) \left(F^N(\theta^k) - F^N(\theta^{k-1}) \right) & k \geq g. \end{cases}$$

Then, the marginal benefit of extending the contest to $T+1$ periods is

$$\begin{aligned}
&EQ^g(\theta^g, N, T+1) - EQ^g(\theta^g, N, T) = \\
&\quad = \sum_{k=1}^{K} \theta^k h^g(\theta^k|\theta^g, N, T+1) - \sum_{k=1}^{K} \theta^k h^g(\theta^k|\theta^g, N, T) \\
&\quad = \sum_{k=1}^{K} \theta^k (h^g(\theta^k|\theta^g, N, T+1) - h^g(\theta^k|\theta^g, N, T)) \\
&\quad = \sum_{k=1}^{g-1} \theta^k \left(F^{NT}(\theta^k)(F^N(\theta^k) - 1) - F^{NT}(\theta^{k-1})(F^N(\theta^{k-1}) - 1) \right) \\
&\qquad + \sum_{k=g}^{K} \theta^k F^{NT}(\theta^{g-1}) \left(F^N(\theta^k) - F^N(\theta^{k-1}) \right)
\end{aligned} \tag{C.6}$$

The principal benefits from extending the contest if

$$F^{TN}(\theta^{g-1}) NC \leq \left(EQ^g(\theta^g, N, T+1) - EQ^g(\theta^g, N, T) \right).$$

From the optimality of θ^g we know that

$$NC \leq \sum_{j=g+1}^{K} \theta^j \left(F^N(\theta^j) - F^N(\theta^{j-1}) \right) - \theta^g \left(1 - F^N(\theta^g) \right)$$

Thus, to show that the agent benefits from extending the contest, it is sufficient to show that

$$\begin{aligned}
F^{TN}(\theta^{g-1}) \Big(\sum_{j=g+1}^{K} \theta^j (F^N(\theta^j) - F^N(\theta^{j-1})) - \theta^g \left(1 - F^N(\theta^g) \right) \Big) \\
\leq \left(EQ^g(\theta^g, N, T+1) - EQ^g(\theta^g, N, T) \right).
\end{aligned} \tag{C.7}$$

Combining (C.6) and (C.7) and simplifying, we get that the sufficient condition is

$$F^{TN}(\theta^{g-1}) \theta^g (1 - F^N(\theta^{g-1})) \geq \sum_{k=1}^{g-1} \theta^k (F^{NT}(\theta^k)(1 - F^N(\theta^k)) - F^{NT}(\theta^{k-1})(1 - F^N(\theta^{k-1}))).$$

Recall that $F(\theta^0) = 0$ and note that we can write this sum as

$$\begin{aligned}
&\sum_{k=1}^{g-1} \theta^k (F^{NT}(\theta^k)(1 - F^N(\theta^k)) - F^{NT}(\theta^{k-1})(1 - F^N(\theta^{k-1})))\\
&\quad = \theta^1 F^{NT}(\theta^1)(1 - F^N(\theta^1)) - \theta^2 F^{NT}(\theta^1)(1 - F^N(\theta^1))\\
&\qquad + \theta^2 F^{NT}(\theta^2)(1 - F^N(\theta^2)) - \theta^3 F^{NT}(\theta^2)(1 - F^N(\theta^2))\\
&\qquad\qquad \vdots\\
&\qquad + \theta^{g-2} F^{NT}(\theta^{g-2})(1 - F^N(\theta^{g-2})) - \theta^{g-1} F^{NT}(\theta^{g-2})(1 - F^N(\theta^{g-2}))\\
&\qquad + \theta^{g-1} F^{NT}(\theta^{g-1})(1 - F^N(\theta^{g-1}))\\
&\quad = -\sum_{k=1}^{g-2} (\theta^{k+1} - \theta^k) F^{NT}(\theta^k)(1 - F^N(\theta^k)) + \theta^{g-1} F^{NT}(\theta^{g-1})(1 - F^N(\theta^{g-1})).
\end{aligned}$$

This allows us to rewrite the sufficient condition to

$$(\theta^g - \theta^{g-1}) F^{TN}(\theta^{g-1})(1 - F^N(\theta^{g-1})) \geq -\sum_{k=1}^{g-2} (\theta^{k+1} - \theta^k) F^{NT}(\theta^k)(1 - F^N(\theta^k))$$

which always holds because $\theta^{k+1} > \theta^k$.

Step 2: To construct an example of a harmful extension of T in case of an FPC we consider a setting with $\delta = 1$, $N = 2$, $\Theta = \{0, 1\}$ and extend $T = 2$ to $T + 1 = 3$. Let the probability of drawing $\theta^1 = 1$ be given by π. We will choose parameters such that the optimal individual threshold is $\theta^i = 1$. The expected costs in the case of $T = 2$ are given by

$$EK(2, 2) = 2(\pi C + 2(1 - \pi)C)$$

and the expected quality

$$EQ(2, 2) = 1 - (1 - \pi)^4.$$

In the case of $T = 3$ we have

$$EK(2, 3) = 2(\pi C + 2(1 - \pi)\pi C + 3(1 - \pi)^2 C)$$

and the expected quality

$$EQ(2, 3) = 1 - (1 - \pi)^6.$$

Hence, the change in expected surplus for the principal is given by

$$EQ(2, 2) - EK(2, 2) - EQ(2, 3) + EK(2, 3) = (1 - \pi)^2 \left(2C - (2 - \pi)\pi(1 - \pi)^2\right)$$

which is negative for $C = 1/10$ and $\pi = 2/5$. Since

$$EQ(2, 2) - EK(2, 2) = \frac{16}{25} - \frac{8}{25} > 0$$

the individual threshold is $\theta^i = 1$ is indeed optimal.

Part IV

Bibliography

Bibliography

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011): "Reference Points and Effort Provision," *American Economic Review*, 101, 470–492.

Apesteguia, J. and M. Ballester (2015): "A Measure of Rationality and Welfare," *Journal of Political Economy*, 123, 1278–1310.

Bartling, B., L. Brandes, and D. Schunk (2015): "Expectations as Reference Points: Field Evidence from Professional Soccer," *Management Science*, 61, 2646–2661.

Bell, D. E. (1985): "Disappointment in Decision Making under Uncertainty," *Operations Research*, 33, 1–27.

Benkert, J.-M., I. Letina, and G. Nöldeke (2017): "On optimal search with infinite horizon," Unpublished manuscript, University of Zurich.

Benzion, U., A. Rapoport, and J. Yagil (1989): "Discount Rates Inferred from Decisions: An Experimental Study," *Management Science*, 35, 270–284.

Bergemann, D. and S. Morris (2005): "Robust Mechanism Design," *Econometrica*, 73, 1771–1813.

Bernheim, B. (2009): "Behavioral Welfare Economics," *Journal of the European Economic Association*, 7, 267–319.

Bernheim, B., A. Popov, and I. Fradkin (2015): "The Welfare Economics of Default Options in 401(k) Plans," *American Economic Review*, 105, 2798–2837.

Bernheim, B. and A. Rangel (2009): "Beyond Revealed Preference: Choice-Theoretic Foundations For Behavioral Welfare Economics," *Quarterly Journal of Economics*, 124, 51–104.

Bierbrauer, F. and N. Netzer (2016): "Mechanism Design and Intentions," *Journal of Economic Theory*, 163, 557–603.

Bimpikis, K., S. Ehsani, and M. Mostagir (2014): "Designing Dynamic Contests," Mimeo.

Border, K. C. (1991): "Implementation of Reduced Form Auctions: A Geometric Approach,"

Econometrica, 59, 1175–1187.

Boudreau, K. J., N. Lacetera, and K. R. Lakhani (2011): "Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis," *Management Science*, 57, 843–863.

Boudreau, K. J., K. R. Lakhani, and M. Menietti (2016): "Performance responses to competition across skill levels in rank-order tournaments: field evidence and implications for tournament design," *The RAND Journal of Economics*, 47, 140–165.

Camerer, C. F., S. Issacharoff, G. Loewenstein, T. O'Donoghue, and M. Rabin (2003): "Regulation for conservatives: behavioral economics and the case for "asymmetric paternalism"," *University of Pennsylvania Law Review*, 151, 1211–1254.

Caplin, A. and D. Martin (2012): "Framing Effects and Optimization," Mimeo.

Carbajal, J. C. and J. C. Ely (2016): "A Model of Price Discrimination under Loss Aversion and State-Contingent Reference Points," *Theoretical Economics*, 11, 455–485.

Chatterjee, K. and W. Samuelson (1983): "Bargaining under Incomplete Information," *Operations Research*, 31, 835–851.

Che, Y.-K. and I. Gale (2003): "Optimal design of research contests," *The American Economic Review*, 93, 646–671.

Che, Y.-K., J. Kim, and K. Mierendorff (2013): "Generalized Reduced-Form Auction: A Network-Flow Approach," *Econometrica*, 81, 2487–2520.

Cohen, J., K. Ericson, D. Laibson, and J. White (2016): "Measuring Time Preferences," Mimeo.

Copic, J. and C. Ponsatí (2008): "Ex-Post Constrained-Efficient Bilateral Trade with Risk-Averse Traders," Mimeo.

Cramton, P., R. Gibbons, and P. Klemperer (1987): "Dissolving a partnership Efficienty," *Econometrica*, 55, 615–632.

Crawford, V. P. (2016): "Efficient Mechanisms for Level-k Bilateral Trading," Mimeo.

Crawford, V. P. and J. Meng (2011): "New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income," *American Economic Review*, 101, 1912–1932.

De Clippel, G. and K. Rozen (2014): "Bounded Rationality and Limited Datasets," Mimeo.

de Meza, D. and D. C. Webb (2007): "Incentive Design under Loss Aversion," *Journal of the European Economic Association*, 5, 66–92.

Dechenaux, E., D. Kovenock, and R. M. Sheremeta (2015): "A survey of Experimental Research on contests, all-pay auctions and tournaments," *Experimental Economics*, 18, 609–669.

DellaVigna, S. (2009): "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, 47, 315–372.

Driesen, B., A. Perea, and H. Peters (2012): "Alternating offers Bargaining with loss aversion," *Mathematical Social Sciences*, 64, 103–118.

Duraj, J. (2015): "Mechanism Design with News Utility," Personal Communcation.

Eisenhuth, R. (2013): "Reference Dependent Mechanism Design," Mimeo.

Eisenhuth, R. and M. Ewers (2012): "Auctions with Loss Averse Bidders," Working paper, Northwestern University.

Ericson, K. M. M. and A. Fuster (2011): "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments," *Quarterly Journal of Economics*, 126, 1879–1907.

——— (2014): "The Endowment Effect," *Annual Review of Economics*, 6, 555–579.

Fehr, E. and L. Goette (2007): "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment," *American Economic Review*, 97, 298–317.

Fieseler, K., T. Kittsteiner, and B. Moldovanu (2003): "Partnerships, lemons, and efficient trade," *Journal of Economic Theory*, 113, 223–234.

Fudenberg, D. and J. Tirole (1991): *Game theory*, Cambridge, Massachusetts: MIT Press.

Fullerton, R. L., B. G. Linster, M. McKee, and S. Slate (2002): "Using Auctions to Reward Tournament Winners: Theory and Experimental Invetigations," *The RAND Journal of Economics*, 33, 62–84.

Gal, S., M. Landsberger, and B. Levykson (1981): "A Compound Strategy for Search in the Labor Market," *International Economic Review*, 22, 597–608.

Garratt, R. and M. Pycia (2015): "Efficient Bilateral Trade," Mimeo, FRBNY and UCLA.

Genesove, D. and C. Mayer (2001): "Loss Aversion and Seller Behavior: Evidence from the Housing Market," *The Quarterly Journal of Economics*, 116, 1233–1260.

Gill, D. and V. Prowse (2012): "A Structural Analysis of Disappointment Aversion in a Real Effort Competition," *American Economic Review*, 102, 469–503.

Gill, D. and R. Stone (2010): "Fairness and desert in tournaments," *Games and Economic Behavior*, 69, 346–364.

Goldin, J. (2015): "Which Way To Nudge? Uncovering Preferences in the Behavioral Age," *Yale Law Journal*, forthcoming.

Goldin, J. and D. Reck (2015): "Preference Identification Under Inconsistent Choice," Mimeo.

Green, B. and C. R. Taylor (2016): "Breakthroughs, Deadlines, and Self-Reported Progress: Contracting for Multistage Projects," *American Economic Review*, 106, 3660–3699.

Grüne-Yanoff, T. (2012): "Old wine in new casks: libertarian paternalism still violates liberal principles," *Social Choice and Welfare*, 38, 635–645.

Halac, M., N. Kartik, and Q. Liu (forthcoming): "Contests for Experimentations," *Journal of Political Economy*.

Hendon, E., H. J. Jacobsen, and B. Sloth (1996): "The One-Shot-Deviation Principle for Sequential Rationality," *Games and Economic Behavior*, 12, 274–282.

Herweg, F., D. Müller, and P. Weinschenk (2010): "Binary Payment Schemes: Moral Hazard and Loss Aversion," *American Economic Review*, 100, 2451–2477.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–291.

Kalai, G., A. Rubinstein, and R. Spiegler (2002): "Rationalizing Choice Functions by Multiple Rationales," *Econometrica*, 70, 2481–2488.

Karle, H., G. Kirchsteiger, and M. Peitz (2015): "Loss Aversion and Consumption Choice: Theory and Experimental Evidence," *American Economic Journal: Microeconomics*, 7, 101–120.

Karle, H. and M. Peitz (2014): "Competition under consumer loss aversion," *The RAND Journal of Economics*, 45, 1–31.

Kőszegi, B. (2014): "Behavioral Contract Theory," *Journal of Economic Literature*, 52, 1075–1118.

Kőszegi, B. and M. Rabin (2006): "A Model of Reference-Dependent Preferences," *The Quarterly Journal of Economics*, 121, 1133–1165.

——— (2007a): "Mistakes in Choice-Based Welfare Analysis," *American Economic Review, Papers and Proceedings*, 97, 477–481.

——— (2007b): "Reference-Dependent Risk Attitudes," *The American Economic Review*, 97, 1047–1073.

——— (2008a): "Choice, Situations, and Happiness," *Journal of Public Economics*, 92, 1821–1832.

——— (2008b): "Revealed Mistakes and Revealed Preferences," in *The Foundations of Positive and Normative Economics*, ed. by A. Caplin and A. Schotter, New York: Oxford University Press, 193–209.

——— (2009): "Reference-Dependent Consumption Plans," *American Economic Review*, 99, 909–936.

Kőszegi, B. and A. Szeidl (2013): "A Model of Focusing in Economic Choice," *Quarterly Journal of Economics*, 128, 53–104.

Konrad, K. A. (2009): *Strategy and Dynamics in Contests*, Oxford University Press.

Kruse, T. and P. Strack (2015): "Optimal stopping with private information," *Journal of Economic Theory*, 159, 702–727.

Kucuksenel, S. (2012): "Behavioral Mechanism Design," *Journal of Public Economic Theory*, 14, 767–789.

Lang, M., C. Seel, and P. Strack (2014): "Deadlines in stochastic contests," *Journal of Mathematical Economics*, 52, 134–142.

Letina, I. (2016): "The Road not Taken: Competition and the R&D Portfolio," *RAND Journal of Economics*, 47, 433–460.

Letina, I. and A. Schmutzler (2016): "Inducing Variety: A Theory of Innovation Contests," Mimeo, University of Zurich.

Loewenstein, G. (1988): "Frames of Mind in Intertemporal Choice," *Management Science*, 34, 200–214.

Loewenstein, G. and D. Prelec (1992): "Anormalies in Intertemporal Choice: Evidence and Interpretation," *Quarterly Journal of Economics*, 107, 573–597.

Loomes, G. and R. Sugden (1986): "Disappointment and Dynamic Consistency in Choice under Uncertainty," *The Review of Economic Studies*, 53, 271–282.

Masatlioglu, Y., D. Nakajima, and E. Ozbay (2012): "Revealed Attention," *American Economic Review*, 102, 2183–2205.

Masatlioglu, Y. and C. Raymond (2016): "A Behavioral Analysis of Stochastic Reference Dependence," *The American Economic Review*, 106, 2760–2782.

Maskin, E. and J. Riley (1984): "Optimal Auctions with Risk Averse Buyers," *Econometrica*, 52, 1473 – 1518.

Mierendorff, K. (2016): "Optimal dynamic mechanism design with deadlines," *Journal of Economic Theory*, 161, 190–222.

Moldovanu, B. and A. Sela (2001): "The optimal allocation of prizes in contests," *American Economic Review*, 542–558.

——— (2006): "Contest architecture," *Journal of Economic Theory*, 126, 70–96.

Morgan, P. B. (1983): "Search and Optimal Sample Size," *The Review of Economic Studies*, 50, 659–675.

Myerson, R. B. and M. A. Satterthwaite (1983): "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory*, 29, 265 – 281.

Paolacci, G., J. Chandler, and P. Ipeirotis (2010): "Running Experiments on Amazon Mechanical Turk," *Judgement and Decision Making*, 5, 411–419.

Pérez-Castrillo, D. and D. Wettstein (2016): "Discrimination in a model of contests with incomplete information about ability," *International Economic Review*, 57, 881–914.

Pope, D. G. and M. E. Schweitzer (2011): "Is Tiger Woods Loss Averse? Persistent Bias

in the Face of Experience, Competition, and High Stakes," *American Economic Review*, 101, 129–157.

Post, T., M. J. van den Assem, G. Baltussen, and R. H. Thaler (2008): "Dear or No Deal? Decision Making under Risk in a Large-Payoff Game Show," *American Economic Review*, 98, 38–71.

Rieck, T. (2010): "Information disclosure in innovation contests," Bonn Econ Discussion Papers 16/2010, Bonn Graduate School of Economics.

Rosato, A. (2014): "Loss Aversion in Sequential Auctions: Endogenous Interdependence, Informational Externalities and the "Afternoon Effect"," Working paper, University of Technology Sydney.

——— (2017): "Sequential Negotiations with Loss-Averse Buyers," *European Economic Review*, 91, 290–304s.

Rubinstein, A. and Y. Salant (2006): "A Model of Choice From Lists," *Theoretical Economics*, 1, 3–17.

——— (2008): "Some Thoughts on the Principle of Revealed Preference," in *Handbooks of Economic Methodologies*, ed. by A. Caplin and A. Schotter, New York: Oxford University Press, 115–124.

——— (2012): "Eliciting Welfare Preferences from Behavioural Data Sets," *Review of Economic Studies*, 79, 375–387.

Salant, Y. and A. Rubinstein (2008): "(A,f): Choice with Frames," *Review of Economic Studies*, 75, 1287–1296.

Salant, Y. and R. Siegel (2016): "Reallocation Costs and Efficiency," *American Economic Journal:Microeconomics*, 8, 203–227.

Schöttner, A. (2008): "Fixed-prize tournaments versus first-price auctions in innovation contests," *Economic Theory*, 35, 57–71.

Shalev, J. (2002): "Loss Aversion and Bargaining," *Theory and Decisions*, 52, 201–232.

Shelley, M. (1993): "Outcome Signs, Question Frames and Discount Rates," *Management Science*, 39, 806–815.

Siegel, R. (2009): "All-Pay Contests," *Econometrica*, 77, 71–92.

Siegel, R. and Y. Salant (2015): "Contracts with Framing," Mimeo.

Spiegler, R. (2012): "Monopoly Pricing when Consumers are Antagonized by Unexpected Price Increases: A "Cover Versio" of the Heidhues-Koszegi-Rabin Model," *Economic Theory*, 51, 695–711.

——— (2015): "On the Equilibrium Effects of Nudging," *Journal of Legal Studies*, 44, 389–416.

Sunstein, C. (2014): *Why Nudge? The Politics of Libertarian Paternalism*, New Haven: Yale University Press.

Taylor, C. R. (1995): "Digging for Golden Carrots: An Analysis of Research Tournaments," *The American Economic Review*, 872–890.

Terwiesch, C. and Y. Xu (2008): "Innovation contests, open innovation, and multiagent problem solving," *Management science*, 54, 1529–1543.

Thaler, R. and C. Sunstein (2003): "Libertarian Paternalism," *American Economic Review, Papers and Proceedings*, 93, 175–179.

——— (2008): *Nudge: Improving Decisions About Health, Wealth, and Happiness*, New Haven: Yale University Press.

Thaler, R. H. (1980): "Toward a positive theory of consumer choice," *Journal of Economic Behavior and Organization*, 1, 39–60.

Weber, E., E. Johnson, K. Milch, H. Chang, J. Brodscholl, and D. Goldstein (2007): "Aymmetric Discounting in Intertemporal Choice – A Query-Theory Account," *Psychologcial Science*, 18, 516–523.

Wolitzky, A. (2016): "Mechanism Design with Maxmin Agents: Theory and an Application to Bilateral Trade," *Theoretical Economics*, 11, 971–1004.

Part V

Curriculum Vitae

Curriculum Vitae

Personal details

Name:	Jean-Michel Benkert
Date of Birth:	March 13, 1988

Education

08/2012 – 07/2017	PhD studies at the Zurich Graduate School of Economics University of Zurich, Switzerland
09/2011 – 07/2012	MSc in Economics and Finance Barcelona Graduate School of Economics, Spain
08/2007 – 07/2010	BA in Business and Economics University of Basel, Switzerland

Professional experience

08/2012 – 02/2017	Research and Teaching Assistant, University of Zurich
09/2011 – 07/2012	Teaching Assistant, Universitat Pompeu Fabra
02/2011 – 07/2011	Junior Research Assistant, University of Basel
08/2010 – 01/2011	Intern Economic Research Switzerland, UBS Investment Bank